Poster at ASSC4:

4th Conference of the Association for the Scientific Study of Consciousness

http://www.ulb.ac.be/event/ASSC4

EVOLVABLE ARCHITECTURES FOR HUMAN-LIKE MINDS

Aaron Sloman, Matthias Scheutz

School of Computer Science, University of Birmingham,

http://www.cs.bham.ac.uk/ {~axs/ ~mxs/}

Brian Logan

School of Computer Science & IT, University of Nottingham

http://www.cs.nott.ac.uk/~bsl/

1 Introduction

Our project combines empirical, philosophical and computational theories and approaches, aiming (eventually) to explain many human, animal and robot phenomena, e.g. varieties of motivation, emotions, perceptual functions, reasoning, learning, and varieties of consciousness.

We regard intuitive concepts of mentality as "cluster" concepts referring to implicitly presupposed virtual machine (VM) architectures.

These pre-theoretical architectures are very crude approximations to diverse actual VM architectures, which vary during development, across species, and after damage or disease.

Many disagreements on topics like 'emotion', 'consciousness', 'intentionality' arise because researchers focus on different sub-clusters of capabilities and mechanisms within such architectures. Researchers studying different aspects of mind are often unaware of what they are ignoring, like the proverbial blind men describing an elephant.

Powerful theories require knowledge about evolution, ethology, neuro-science, psychology, phenomenology, and feasibility results from AI.

Our draft 'Cogaff' architecture-schema for human-like minds models *reactive*, *deliberative* and *meta-management* layers, which evolved at

different times. The layers operate concurrently and asynchronously, interacting in a variety of ways, producing varieties of emotions, learning, perception, motivation, etc.

Human-like self-consciousness requires the third layer, though more primitive sorts are implemented in older layers. Future work includes understanding many discontinuities and trade-offs in the evolution of mental architectures.

Variations in the architecture should accommodate infants and adults, normal and pathological cases, and also animals.

These architectures provide a framework for systematically generating families of concepts of various kinds of mental phenomena. Such concepts refine and extend familiar concepts of mind, providing a new, richer, more precise theory-based ontology. Within this unifying framework we hope to explain and resolve disagreements and move towards deeper explanatory theories and more powerful and realistic artificial models, for use in many applications, including education and entertainment.

2 Approaches to the study of mind

There are many approaches to the study of mind, and much ambiguity in words like 'emotion' and 'consciousness'. E.g. Dennett recommends the *intentional stance* as accounting for our ascription of mental concepts. By contrast, we view mentality as concerned with an *information level* architecture, close to the level of description used by software engineers. This extends the *design stance* by using a level of description between physical levels and intentional descriptions, attempting to understand human minds as information processing virtual machines with a complex multi-level architecture, whose components evolved at different times and perform different sorts of functions.

There are many types of virtual machines, some implemented in digital computers, some in mixtures of digital and analog circuitry, some in brains. Examples include software virtual machines, minds, social and economic processes. All of these are ultimately implemented in physics, though we don't yet know how, for most of them.

There are several types of *computational* approaches to mind, which produce theories and models of varying depth. Some theorists strive for

realism by modelling known neural structures and processes. Others aim for a more abstract level of modelling.

We explore VM architectures at different levels of abstraction, expecting later work to move further ‘downwards’ towards implementations that are closer to biological details and also ‘upwards’ by showing how the architecture can explain a wide range of phenomena including social phenomena.

3 Constraining the architecture

Indefinitely many VM architectures could in principle produce any observed behaviours of an organism. How can we search for *correct* explanations? There is no guaranteed method. But we can be guided by:

(1) *Information about the physical design of the system.*

(2) *Information about the design history, e.g. its evolution.*

(3) *Observed behaviour in ‘natural’ and laboratory contexts.*

(4) *Introspection: used with caution*.

(5) *Broken parts: effects of brain damage, disease, etc.*.

(6) *Explicit or implicit common sense knowledge: a* vast *source of ideas.*

(7) *Information from several other disciplines*.

(8) *Requirements for unified explanations of diverse phenomena.*

(9) *Analogies with better understood systems, e.g. designed systems*.

(10) *What we have learnt in AI and software engineering.*

(11) *Careful comparisons between competing theories*.

4 Virtual machines are not epiphenomenal

A VM may be fully implemented in physical systems yet be real and have effects. E.g. economic inflation can cause poverty and poverty can cause ignorance and crime. These can have physical effects, e.g. motion of stolen TV sets.

Likewise thoughts, feelings and decisions can cause thoughts feelings and decisions, and also cause physical behaviour.

But this does not require physics to be causally incomplete: causation

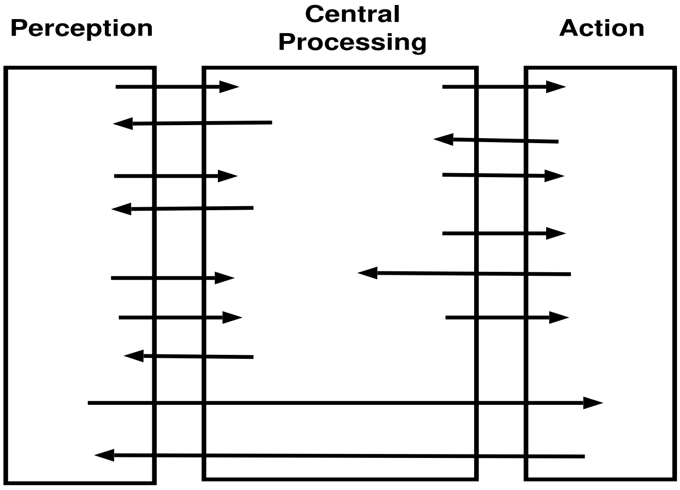

Figure 1: Towers with horizontal information flow

is not some kind of conserved fluid. Rather it is a matter of the truth of very complex sets of conditional statements. These do not have to be statements about physical phenomena.

We do not yet know how many types of virtual machines there are, how they can be implemented, what their powers are. Familiarity with theorems about Turing machines, or properties of digital computers does not guarantee any deep understanding of the space of types of computational virtual machines, including machines with large numbers of coexisting, concurrent asynchronously interacting components, e.g. animal and robot minds.

'Virtual machine' is now a familiar notion, but there is still work to be done analysing this and related notions: *supervenience*, *implementation* and *realisation*. Philosophers are mostly unaware of relevant insights from software engineering.

5 Architectural decomposition

We can superimpose two commonly used architectural decompositions, one 'vertical' (towers) and one 'horizontal' (layers) illustrated below.

Figure 1 shows information flow through the sensory, central and motor sub-systems. Each 'tower' may comprise simple transducers or may have sophisticated multi-functional layered mechanisms, e.g. perceptual processes combining segmentation, recognition, interpretation, attention control, and sub-mechanisms interacting through intermediate information stores recording partial results. (Sloman, 1989). Likewise action sub-systems may have complex hierarchical mechanisms, and can vary in kinds of feedback and degree of coordination with perceptual or central systems. (Albus, 1981 Ch 7).

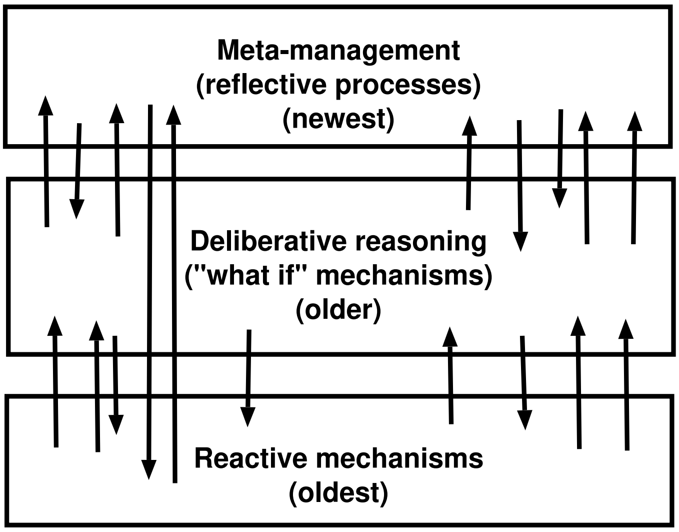

Figure 2: Levels of control and sophistication

Figure 2 depicts processing layers which evolved at different times, performing tasks of varying abstractness. Processing in different layers may be concurrent, asynchronous, and either cooperative or mutually disruptive! In some models higher levels dominate lower levels. In ours, higher levels mainly perform more sophisticated functions.

A three tower model could refer to the physical architecture, or to abstract virtual machine functionality. In the latter case, distinct VM sub-systems could be implemented in *shared* physical mechanisms. E.g. vision and spatial reasoning might share some brain mechanisms.

5.1 *Varieties of layers*

‘Layered’ models propose a variety of forms of information processing. Our Cogaff model (partly inspired by Simon (1967)) postulates three main concurrently active layers which evolved at different times. They account for processes in different kinds of animals and provide a framework for distinguishing different concepts of ‘emotion’.

The first layer contains *reactive* mechanisms which automatically take action as soon as appropriate conditions are satisfied. The second *deliberative* layer provides ‘what if’ reasoning capabilities, required for planning, predicting and explaining. Relatively few organisms have this, and again the forms can vary widely. The *meta-management* layer

provides the ability to monitor, evaluate, and partly control, internal processes and strategies.

Deliberative mechanisms can contemplate possible novel action sequences, and situations, compare them, evaluate them and select them. In humans, they can also consider hypothetical past or future situations. They are inherently slower, more discrete and more sequential than reactive mechanisms (Sloman 2000).

A meta-management system can monitor, categorise, evaluate, and sometimes redirect some (not all) internal processes. None of the three layers has total control: they can interrupt one another and all can be disrupted by global alarms (described below).

Various reactive mechanisms occur in all organisms, and are very old. Deliberative mechanisms evolved later. Can a rat or chimp wonder what would happen if...? (Cf Kohler 1927.) Meta-management mechanisms evolved most recently. Perhaps chimpanzees have simple versions?

Manipulating goals, planning, tailoring actions to contexts, recognising *affordances* (Gibson 1979), all require various representational mechanisms. Deliberative requirements induce a need both for more abstract representational capabilities and also more sophisticated perceptual processing to do chunking and generalisation in terms of features of the environment, not the sensory array. These operate in parallel with more 'primitive' sensory processing e.g. for posture control.

Of course, all the different layers must ultimately be implemented in purely reactive mechanisms, otherwise nothing would ever happen.

5.2 *How did the layers evolve?*

Biological evolution often produces two copies of an old structure, one of which then develops a new function. Maynard Smith & Szathmáry call this 'duplication and divergence'. E.g. Reactive condition-action mechanisms might be copied and modified to form a long term associative memory needed for 'what-if' reasoning.

Elsewhere we argue that the diversity of requirements of central components provides evolutionary pressure for functional differentiation of both perceptual sub-mechanisms and action subsystems. This in turn provides opportunities for further development and differentiation of central functions. This leads to a view of mind as a co-evolved *eco-system* of sub-organisms (Figure 3).

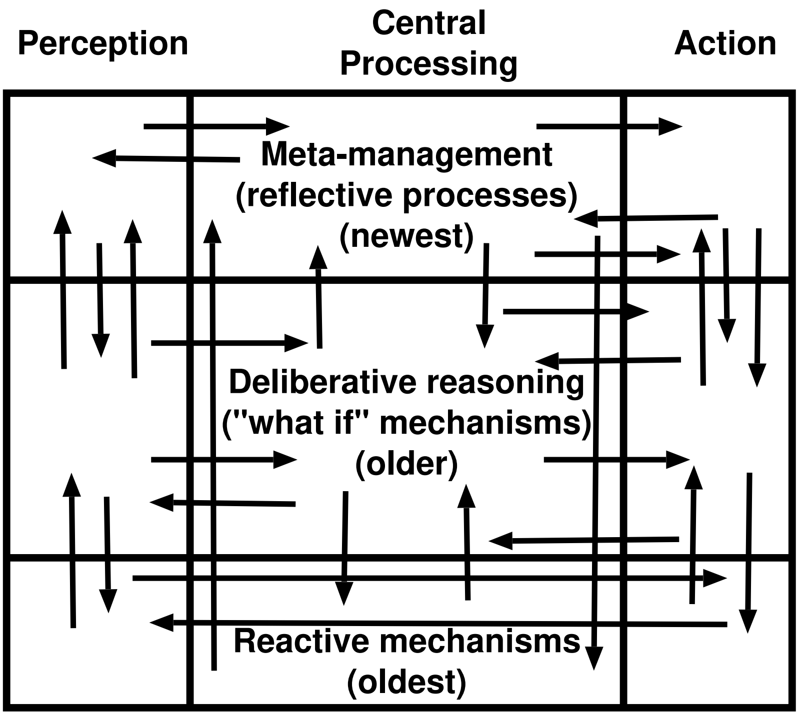

Figure 3: A grid of co-evolved interacting sub-modules

5.3 *Executive function?*

Researchers who refer to 'executive functions' are often unaware that they are discussing mechanisms which combine deliberative and meta-management, which can come apart, e.g. after brain damage (cf. Damasio 1994). Some AI planning systems do not have meta-management layers, and may not detect that their reasoning strategies are wasteful or lead to loops.

6 Further developments: grids and alarms

The grid in Figure 3 superimposes the towers and layers. Figure 4 adds a reactive *global alarm* mechanism that can detect situations where urgent action is needed to take an opportunity or avoid a disaster.

An alarm mechanism might receive inputs from many parts of the system, and send out overriding control signals. Using fast general pattern-recognition techniques it can interrupt everything and produce a new global response, e.g. fleeing, freezing, pouncing, attending, etc. These ideas were partly inspired by engineering requirements (Simon, 1967), partly by studies of the brain, e.g. the limbic system (Albus, 1981; LeDoux, 1996).

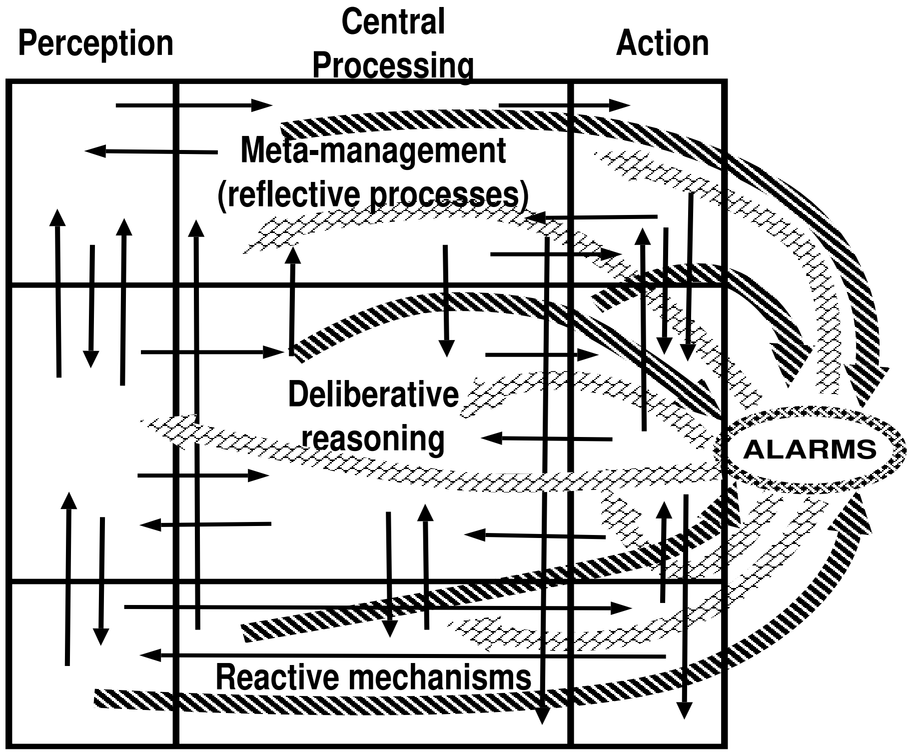

Figure 4: **Adding one or more fast but 'stupid' globally connected alarm systems.**

Alarm mechanisms need to react quickly, so they will be pattern-based, using both innate and learnt patterns and responses. Being fast and stupid, they are error prone, but may be improved by training.

Figure 5 lists additional components required for processing of motives, 'what if' reasoning capabilities in the deliberative layer, and aspects of self-control. E.g. an attention filter can compensate for limited deliberative resources.

Different, culturally influenced, 'personae' (global processing patterns) could be active in different contexts, e.g. when at home with family, when driving a car, when with subordinates in the office, in the pub with friends, etc.

Figure 6 assembles many of the pieces, impressionistically portraying a schema for a human-like architecture (adult, not infant).

Simpler architectures e.g. containing only reactive mechanisms with alarms, may suffice for other organisms, such as insects.

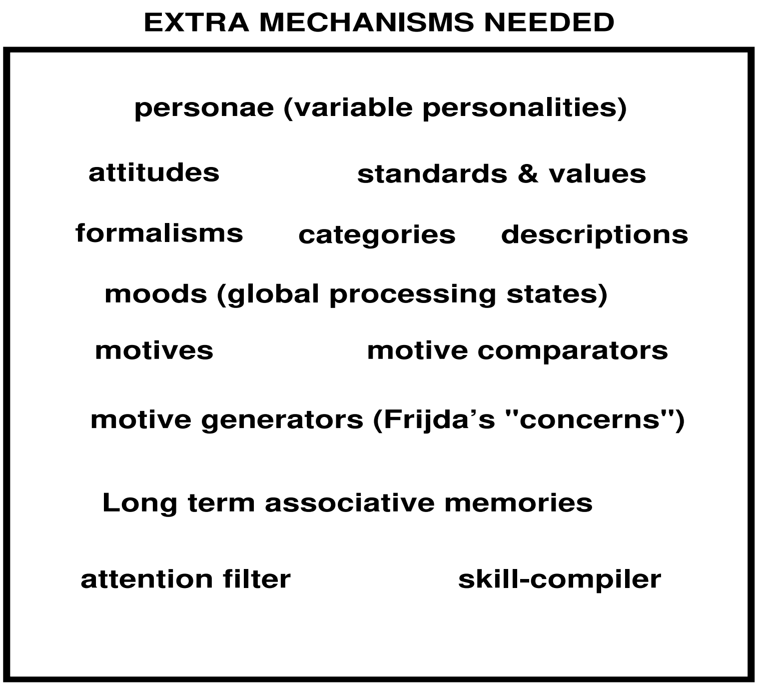

Figure 5: Additional components required for the top two layers

7 Motive generators, comparators, evaluators

Reactive systems may use implicit or explicit goals. Deliberative systems need *explicit* goals to drive planning and access relevant generalisations from a long term memory.

If not all motives are permanently present, *motive generators* can create new goals. (Beaudoin 1994). Compare Frijda's dispositional 'concerns'. Motive *comparators* are useful for dealing with conflicts.

Separate evaluation mechanisms are also needed for assessing current and future states.

8 Dynamic filters and moods

Under stressful conditions, deliberative mechanisms with limited resources can become overloaded by interrupts. Variable-threshold attention filters can reduce this problem (Beaudoin & Sloman 1993). Malfunctions could produce attention disorders (Beaudoin 1994).

More coarse-grained global control mechanisms producing enduring changes may account for some *moods*. In a 'friendly' environment where most goals are achieved, a confident mood may be fruitful. When things often go wrong, a more cautious, pessimistic mood may be safer. More subtle and complex changes of mood may be triggered socially.

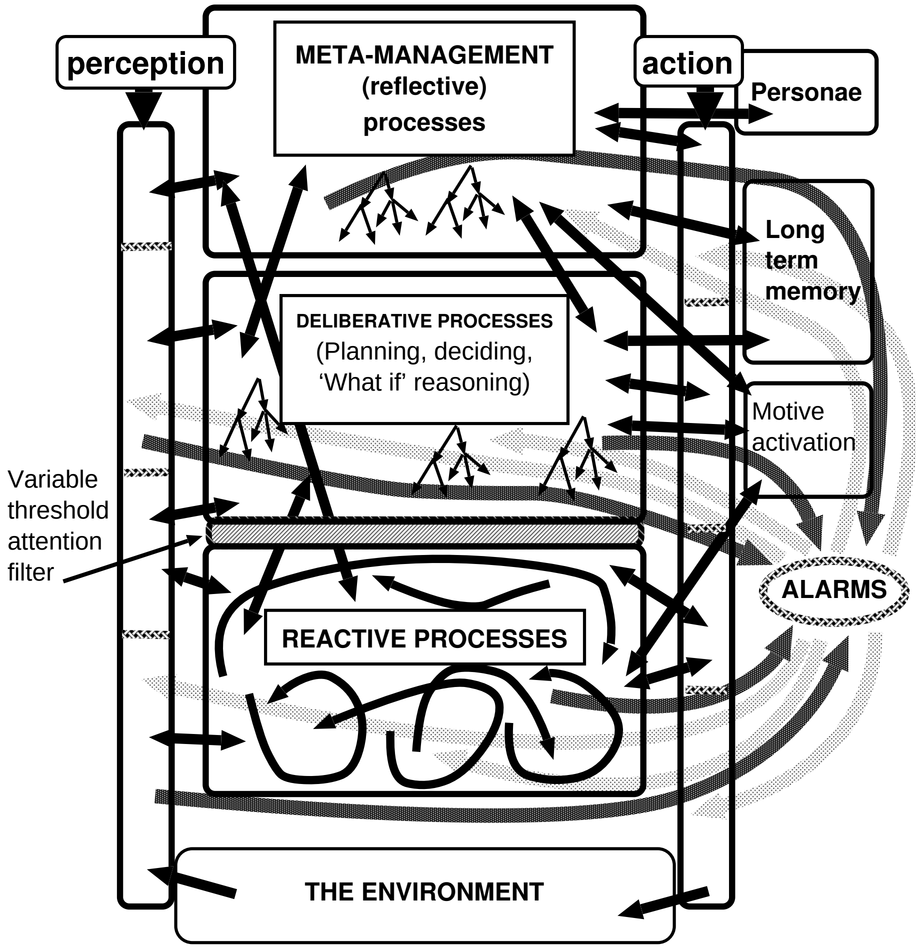

Figure 6: The Birmingham *'Cogaff'* architecture-schema

9 Related theories

Theories with layered architectures are quite old, e.g. Freud's *id*, *ego* and *super-ego*. Albus (1981, page 184) presents the notion of a layered 'triune' brain with a reptilian lowest level and old and new mammalian levels. There are many variants, e.g. Brooks' (1991) 'subsumption hierarchy', Minsky (1987), Nilsson (1998) and Johnson-Laird's discussion (1993) of consciousness.

In some theories, it is proposed that as sensory information comes in, increasingly abstract interpretations flow up through various layers. At the 'top' they may trigger decision-making which causes instructions to trickle down through the layers and out through the motors. We call this the 'Omega' model since the flow pattern looks a bit like an Ω.

In our Cogaff model all layers get information (of different degrees of abstraction) in parallel and process it in parallel and produce action

signals (of different degrees of abstraction) in parallel, e.g. simultaneously walking and discussing philosophy, while digesting food, controlling posture, enjoying the view, etc. The deliberative system cannot directly control reactive mechanisms, but may be able to *train* some of them (e.g. learning skills through practice).

10 Architecture-based concepts

An architecture determines a collection of mental concepts grounded in the varieties of states and processes it can support. Architecture-based concepts may prove superior to those we currently use to formulate research questions and theories.

We often think we know exactly what consciousness, experience, emotions, etc. are, because we experience them directly. This is mistaken. We may experience *simultaneity* 'directly' sometimes, but that does not guarantee a clear grasp of the concept, as Einstein showed. Architecture-based concepts can refine and extend old cluster concepts without replacing them.

11 Architecture-based emotion concepts

Previous work showed how the architecture supports processes corresponding to some pre-scientific concepts of 'emotion'. The architecture supports the common distinction between 'primary' and 'secondary' emotions (e.g. Damasio, 1994; Picard, 1997) and suggests an additional category of 'tertiary' emotions (and possibly more).

Primary emotions (e.g. being startled) mainly involve the reactive layer. The deliberative layer explains semantically rich, centrally triggered *secondary* emotions of various kinds, e.g. apprehension. Types of partial loss of control of attention in the meta-management layer account for *tertiary* emotions, involving partial loss of control of attention (e.g. in infatuation). Tertiary emotions are probably unique to humans, though perhaps simplified versions can be found in some other animals.

There are myriad definitions of 'emotion'. Further work on architecture-based concepts may account for disagreements in definition and generate a host of new more refined concepts.

12 Multiple personalities

The style of meta-management can vary. Different global patterns of organisation, i.e. different ‘personae’, or personalities, are appropriate in different contexts. Switching to a different personality involves turning on a large collection of skills, styles of thought and action, types of evaluations, etc.

This requires (i) a store of ‘personae’, (ii) mechanisms for acquiring new ones, (iii) mechanisms for modifying or extending them, (iv) mechanisms for ‘switching control’ between them. This is a powerful method of social control, saving evolution a lot of work. It can also be dysfunctional, e.g. in religious indoctrination. Malfunctions in such systems might produce some types of multiple personality disorders.

13 An eco-system of mind

During evolution, changing needs of central components might cause evolution of new perceptual and motor mechanisms, and vice versa.

We can think of the system as forming an eco-system of co-evolving sub-organisms, with complex parallel trajectories through both ‘design space’ and ‘niche space’ (Sloman 1998).

Co-evolution within an organism is more likely to be cooperative than competitive, though there could also be competition, e.g. for information, blood supplies, etc. (Maynard Smith & Szathmáry Ch 8).

Although the notion of an ‘eco-system’ must not be taken too literally, nevertheless, trying to understand processes of incremental changes in different parts of the architecture may help us understand how the whole system evolved, and how that system works. (Popper 1976, p. 173)

14 Conjecture: qualia

The three layers can explain different notions of consciousness, ranging from simple sentience (e.g. in insects) to full reflective self-awareness and possession of ‘qualia’ (e.g. attendable intermediate contents of sensory processing (Sloman 2000)). Robots with appropriate architectures would have such qualia and be aware of them.

15 Conclusion

An architecture-based ontology can bring some order into the morass of studies of affect and consciousness, explaining how many conflicting theories and inconsistent definitions arise out of attending to different subsets of a huge array of phenomena. We have begun to illustrate this though there is still much work to be done, e.g. linking up to empirical results reported by Baars and many others.

E.g. The Cogaff model predicts (Sloman 1989) that there will be many visual pathways for different sub-tasks, operating at different speeds, and using sub-mechanisms of varying sophistication in the eco-system of mind. Learnt skills (e.g. musical sight-reading, tennis) can produce new pathways. Compare (Goodale & Milner 1992). There may be many undiscovered functionally distinct *virtual machine* pathways.

Developmental studies are also important. Human infants are probably not born with a fully formed architecture, though simpler organisms may have innately determined architectures. The Cogaff architecture suggests a rich variety of types of individual development and learning.

All this has practical implications for engineers and scientists who design working systems that:
(a) model human vision, affect, learning, language understanding, etc.
(b) model 'believable' synthetic agents in interactive entertainments,
(c) function as interactive teaching/training packages for would-be counsellors, psychotherapists, psychologists, brain scientists.

16 Acknowledgements & Notes

This work is partly funded by the Leverhulme trust, and has benefitted from interactions with many people at Birmingham and elsewhere.

Project papers can be found at:

`http://www.cs.bham.ac.uk/research/cogaff/`

Our software tools are available free of charge with full system sources at the Free Poplog site:

`http://www.cs.bham.ac.uk/research/poplog/freepoplog.html`

References

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

Baars, B. J. (1988), *A cognitive Theory of Consciousness*, Cambridge University Press, Cambridge, UK.

Beaudoin, L. (1994), Goal processing in autonomous agents, PhD thesis, School of Computer Science, The University of Birmingham. (Available at http://www.cs.bham.ac.uk/ research/ cogaff/).

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

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

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

Frijda, N. H. (1986), *The emotions*, Cambridge University Press, Cambridge.

Goodale, M. & Milner, A. (1992), 'Separate visual pathways for perception and action', *Trends in Neurosciences* 15(1), 20–25.

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

Kohler, W. (1927), *The Mentality Of Apes*, Routledge & Kegan Paul, London. 2nd edition.

LeDoux, J. E. (1996), *The Emotional Brain*, Simon & Schuster, New York.

Maynard Smith, J. & Szathmaŕy, E. (1999), *The Origins of Life: From the Birth of Life to the Origin of Language*, Oxford University Press, Oxford.

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

Simon, H. A. (1967), 'Motivational and emotional controls of cognition'. Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979.

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.

Sloman, A. (1998), Damasio, Descartes, alarms and meta-management, *in* 'Proceedings International Conference on Systems, Man, and Cybernetics (SMC98), San Diego', IEEE, pp. 2652–7.

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

Wright, I., Sloman, A. & Beaudoin, L. (1996), 'Towards a design-based analysis of emotional episodes', *Philosophy Psychiatry and Psychology* 3(2), 101–126.