Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cognitive Systems Research xxx (2004) xxx–xxx
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

More things than are dreamt of in your biology: Information-processing in biologically inspired robots ☆

Action editor: Ron Sun

A. Sloman [a,*], R.L. Chrisley [b]

[a] *School of Computer Science, University of Birmingham, UK*
[b] *Centre for Research in Cognitive Science, University of Sussex, UK*

Received 22 February 2003; accepted 7 June 2004

Abstract

Animals and robots perceiving and acting in a world require an ontology that accommodates entities, processes, states of affairs, etc., in their environment. If the perceived environment includes information-processing systems, the ontology should reflect that. Scientists studying such systems need an ontology that includes the first-order ontology characterising physical phenomena, the second-order ontology characterising perceivers of physical phenomena, and a (recursive) third order ontology characterising perceivers of perceivers, including introspectors. We argue that second- and third-order ontologies refer to contents of *virtual* machines and examine requirements for scientific investigation of combined virtual and physical machines, such as animals and robots. We show how the CogAff architecture schema, combining reactive, deliberative, and meta-management categories, provides a first draft schematic third-order ontology for describing a wide range of natural and artificial agents. Many previously proposed architectures use only a subset of CogAff, including subsumption architectures, contention-scheduling systems, architectures with 'executive functions' and a variety of types of 'Omega' architectures. Adding a multiply-connected, fast-acting 'alarm' mechanism within the CogAff framework accounts for several varieties of emotions. H-CogAff, a special case of CogAff, is postulated as a minimal architecture specification for a human-like system. We illustrate use of the CogAff schema in comparing H-CogAff with Clarion, a well known architecture. One implication is that reliance on concepts tied to observation and experiment can harmfully restrict explanatory theorising, since what an information processor is doing cannot, in general, be determined by using the standard observational techniques of the physical sciences or laboratory experiments. Like theoretical physics, cognitive science needs to be highly speculative to make progress.

☆ Revised version of a paper presented at the WGW'02 workshop on *Biologically inspired robotics: The legacy of W. Grey Walter*, Hewlett–Packard Research Labs, Bristol, August 2002.

* Corresponding author.
E-mail addresses: A.Sloman@cs.bham.ac.uk (A. Sloman), R.L.Chrisley@sussex.ac.uk (R.L. Chrisley).
URLs: http://www.cs.bham.ac.uk/~axs/, http://www.cogs. susx.ac.uk/users/ronc/.

1389-0417/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.cogsys.2004.06.004

Keywords: Architecture; Biology; Emotion; Evolution; Information-processing; Ontology; Ontological blindness; Robotics; Virtual machines

1. Ontologies and information processing

An ontology used by an organism or robot is the set of objects, properties, processes, etc. that the organism (be it a scientist or a seagull) or robot recognises, thinks in terms of, and refers to in its interactions with the world. This paper discusses some of the components of an ontology required both for an understanding of biological phenomena and for the design of biologically inspired robots. The ontology used by scientists and engineers studying organisms and designing robots will have to include reference to the mechanisms, forms of representation and information-processing architectures of the organisms or robots. Insofar as these natural or artificial agents process information, they will use ontologies. So the ontologies used by scientists and engineers will have to refer to those ontologies. That is, they will have to include meta-ontologies. If we wish to talk about many different organisms or robots (e.g., in discussing evolution, comparing different animals in an ecosystem, or comparing robot designs) our ontology will need to encompass a variety of architectures. At present such comparative studies are hampered by the fact that different authors use different terminology in their ontologies, and produce architecture diagrams using different conventions that make it difficult to make comparisons. In this paper, we present an approach to developing a common framework for describing and comparing animals and robots, by introducing a schematic ontology for some of the high level aspects of a design. We do not claim that this is adequate for all the systems studied in AI, psychology and ethology, but offer it as a first step, to be refined and extended over time.

1.1. Non-physical aspects of organisms and their environments

It is relatively easy to observe the gross physical behaviour of organisms, their physical environment, and to some extent, their internal physical, chemical, physiological mechanisms. But insofar as biological organisms are to a large extent control systems (Wiener, 1961), or more generally *information-processing* systems, finding out what they do as controllers or as information processors is a very different task from observing physical behaviour, whether internal or external (Sloman, 1993, 2003). [1]

That is because the most important components of an information processor may be components of *virtual* machines rather than *physical* machines. Like physical machines, virtual machines do what they do by virtue of the causal interaction of their parts, but such parts are non-physical (by 'non-physical', we do not mean 'not physically realised' or 'made ultimately of non-physical stuff' but merely 'not easily characterised with the vocabulary and methods of the physical sciences'). Compare the notion of a 'propaganda machine'. Entities in virtual machines can include such things as grammars, parsers, decision makers, motive generators, inference engines, knowledge stores, recursive data-structures, rule sets, concepts, plans

[1] Throughout this paper, we use 'information' in the colloquial sense in which information is *about* something rather than in the technical sense of Shannon. That is, like many biologists, software engineers, news reporters, information agencies and social scientists, we use 'information' in the sense in which information can be true or false, or can more or less accurately fit some situation, and in which one item of information can be inconsistent with another, or can be derived from another, or may be more general or more specific than another. None of this implies that the information is expressed or encoded in any particular form, such as sentences or pictures or neural states, or that it is communicated between organisms, as opposed to being acquired or used by one organism. We have no space to rebut the argument in (Rose, 1993) that only computers, not animals or brains, are information processors, and the 'opposite' argument of Maturana and Varela summarised in (Boden, 2000) according to which *only* humans process information, namely when they communicate via external messages.

and emotional states, rather than molecules, transistors or neurones.

An example of a component of a virtual machine in biology is the niche of a species. A niche is not a geographical location or a physical environment; for an ant, a badger, and a cat may be in the same physical location yet have very different niches, providing different information for them to process, e.g., different affordances such as opportunities, threats and obstacles (Gibson, 1986).

The niche is not something that can be observed or measured using instruments employed in the physical sciences. Yet the niche is causally very important, both in the way that the organism works (e.g., as an information processor) and in the way that a species evolves (Sloman, 2000a). A niche is part of what determines features of new generations, and in some cases may be involved in reproducing itself also, for instance if members of a species all alter the environment in such a way as to enhance their biological fitness. An example would be termites building and maintaining their cathedrals, which help to produce new generations which will do the same. So the niche, the set of abstract properties common to results of such genetically induced actions, could be labelled as part of an 'extended genotype', by analogy with Dawkins' "extended phenotype" (Dawkins, 1982).

Additional conceptual problems bedevil the task of deciding what features, especially non-physical ones, of a biological system are to be replicated in robots. For instance, many of our colloquial concepts are inadequate for specifying design features. For example, the question whether a certain animal, or robot, has emotions or is conscious or feels pain suffers from the multiple confusions in our current notion(s) of mental states and processes (Sloman, 2002a, 2001a, Sloman, Chrisley, & Scheutz, 2004). So, in part, our task is to explain how to make those obscure concepts clearer, for instance by interpreting them as 'architecture-based' concepts (Sloman & Chrisley, 2003). [2]

[2] In (Sloman & Chrisley, 2003), we contrast 'architecture-based concepts', used in referring to systems with a particular sort of architecture, and 'architecture-driven concepts' used by organisms or robots with a particular architecture, and show how certain architectures may support the use of architecture-driven concepts referring to qualia.

1.2. Orders of ontology

The fact that all organisms acquire and use information, and some also store it, transform it, derive new information from old, and combine it in various ways, places strong constraints on the ontology appropriate for a scientific understanding of organisms, or the ontology used in designing biologically-inspired robots.

Obviously, organisms are also physical systems, which can be described using the ontology of the physical sciences (physics and chemistry). But it has long been recognized that an extended ontology based on a notion of information is useful in biology. Although talk of information processing by organisms (and by low-level components of organisms, such as neurons) is now commonplace in biology, there remains the task of finding out exactly *what* information is acquired, used or derived by particular sorts of organisms, and also how it is represented and what mechanisms manipulate it.

Any system which processes information will have its own ontology: the objects, properties, processes etc. that the information that the system processes is about. In some cases it will be about information processing, whether in itself or in something else. Therefore, we can make a distinction between different orders of ontology required for describing natural systems and designing biologically inspired robots. A *first-order* ontology is an ontology used to refer to arbitrary entities, properties, relationships, processes, etc., for instance an ontology including physical objects, properties such as mass, length, chemical concentrations, and so on. But the designer or scientist may wish to refer to something that includes information-processing, representations, perception, etc. In that case, a subset of the designer's ontology will be a *second-order* ontology: which refers to another ontology used by the system (organism or robot) under consideration.

Furthermore, some organisms (and some robots) also have to take account of the fact that some of the entities in their environment are information-processors, or that they themselves are. These organisms will somehow need to use an appropriate ontology to enable them to make use of information about information-processing sys-

tems. So if one animal (or robot), A, takes account of what another animal (or robot), B, perceives, wants, intends, knows, etc., then part of A's ontology includes a *second-order* ontology. The scientist or designer who talks about A's ontology will be using a *third-order* ontology in that case. The ontology used by A need not have the generality of theoretical computer science, cybernetics or philosophy, but will be suited to the organism's or robot's own needs and its capabilities, which can vary enormously, both between individual organisms and within the lifetime of one organism. All but the first-order ontologies involve semantic content, referring to entities with semantic contents (e.g., plans, percepts, intentions, etc.). We therefore label them as *meta-semantic* ontologies, a notion that will be seen to be important in the discussion of architectures with meta-management, below. Obviously, ontologies can continue to nest to arbitrarily higher orders, but these three orders of ontology should suffice for the points we wish to make.

The requirements on depth, precision and correctness of an ontology will vary, depending on who is using the ontology and for what purposes. The third-order ontology used by a scientist or engineer to talk about A will need considerable clarity and precision, even though the second-order ontology used by A to think about B falls far short of that, since A is not designing B or explaining how B works. Human designers and scientists often switch between using second-order ontologies that are adequate for ordinary life (e.g., talking about emotions of other people) and using third-order ontologies without realising that the concepts in their second-order ontologies will not suffice for use in scientific third-order ontologies (Sloman et al., 2004).

2. Ontologies in science, and how they change

Progress in science takes many forms, including discovering generalisations, refuting generalisations, and discovering new observable types of phenomena. Many of those are discoveries use an existing ontology. If your ontology already includes pressure, volume and temperature as properties of a sample of gas, then no new entities need be postulated in order to formulate laws relating variations in pressure, volume and temperature. [3]

Sometimes scientific progress requires a change in ontology. For example, the discovery that gases are made of previously unknown particles with new kinds of properties (e.g., molecules with mutually repulsive forces) required an extension of the ontology of physics to accommodate the new entities and their properties. In general the deepest advances (both in science and in the conceptual development of an individual) are those that extend our ontologies – for they open up both new classes of questions to pose and new forms of explanations to be investigated. These are not cases where the ontology can be extended simply by *defining* new concepts in terms of old ones: far more subtle and complex processes are involved, as explained in (Sloman, 1978, chap. 2) and in (Carnap, 1947). [4]

2.1. Multi-level ontologies

Some extensions to an ontology are simple additions, for instance adding a new sub-atomic particle or a new type of force to the ontology of physics. Others involve creation of a new ontological level with its own objects, properties, relations, events and processes. Sometimes a new ontological level is proposed as lying 'below' the previous levels and providing a deeper explanation for them (as happened when sub-atomic particles were added to physics, and more profoundly when quantum mechanics was added to physics). It is also possible to propose a new 'higher' ontological level whose entities and processes are somehow based on or dependent on a previously known 'lower' ontological level. An example is the ontological level of biology, including notions like

[3] However, as Newton and many other scientists have discovered, a mathematical ontology may need to be extended. For instance, people may understand that changes can be measured but lack the concept of an instantaneous velocity, or may know about velocity but be unable to think about acceleration.

[4] A partial critique of the idea of 'Symbol grounding' as a solution to this problem is presented in http://www.cs.bham.ac.uk/research/cogaff/talks/#talk14.

gene, species, inheritance, fitness and niche, all of which are nowadays assumed to be in some sense based on the ontological level of physics, though the precise relationship is a matter of debate. A less well-known case is the 'autopoesis' ontological level associated with Maturana and Varela involving notions of self-organisation, self-maintenance, self-repair, etc. discussed in (Boden, 2000). An even more controversial case is the Gaia ontological level proposed in (Lovelock, 1979), scorned by some scientists but not all.

The use of higher ontological levels is not a peculiarity of science: our ordinary mental and social life would not be possible without the use of ontologies involving mental, social, economic, legal, and political entities, properties, relationships, events and processes. For example, our society makes heavy use of interlinked notions of law, transgression, punishment, motive, belief, decision, etc. It seems that some other animals may have simplified versions of such ontological levels, insofar as they acquire and use information about dominance hierarchies, for instance. Since these things, like the niches mentioned previously, can be involved in causal relationships (e.g., ignorance can cause poverty, and poverty can sometimes cause crime) we can think of them as parts of a *machine*, namely a 'virtual machine', in the sense in which a running computer operating system is a virtual machine. This is explained below.

In general the relationships between ontological levels are not well understood: we use intuitive, informal ways of thinking about them, though these can generate apparently irresolvable disputes, for instance disputes about whether progress in science should eliminate or justify the ontological level of folk-psychology.

2.2. Virtual machine ontologies

A species of ontological layering that is easier to understand than most is found in computing systems where the ontological level of a *virtual machine* (e.g., a chess-playing machine, a compiler, a theorem prover, an operating system) is implemented on top of an underlying *digital electronic machine*, a relation often mediated by a hierarchy of intermediate virtual machines. Unlike most other cases, the ontological level of software virtual machines in computers is a product of human design. Consequently, insofar as we have designed and implemented these machines, and know how to modify, extend, debug, and use them, we have a fairly deep understanding of what they are and how they work, though this case is generally ignored by most philosophers and scientists discussing ontological levels and supervenience (e.g., Kim, 1998).

Articulating and formalising all the features of natural or artificial information-processing systems poses many difficulties, including the difficulty of analysing the causal powers of virtual machine events and processes, discussed in more detail in (Sloman & Scheutz, 2001). [5]

For those who study animals or design robots there is a further complication, namely, that the subject of investigation is an information-processing system that must itself (implicitly or explicitly) use an ontology which delimits the types of things it can perceive, think about, learn, desire, decide, etc. Moreover, in more sophisticated cases the information-processing architecture can, as in humans, extend the ontology it uses. It follows that whereas most scientists (e.g., physicists, chemists, geologists) can use ontologies without thinking about them or understanding how they work, this is a luxury that roboticists and biologists cannot afford if we wish to understand animal behaviour or design intelligent robots. Roboticists who successfully design and implement information-processing virtual machines forming control systems for their robots must have at least an intuitive grasp of ontological layering, in contrast with those who eschew design in favour of evolving robot controllers. It is possible to produce artificially evolved systems that are as little understood as products of biological evolution.

Scientists and engineers need to understand the variety of processes by which deployed ontologies develop. We previously noted that there is a kind of intelligence and problem-solving that involves

[5] The discussion is extended in http://www.cs.bham.ac.uk/research/cogaff/talks/ in talks 22, 23 and 26.

the development of new ontologies, of which certain forms of scientific advance are an important special case. Most forms of ontological change have not been modelled in AI or explained theoretically. If we wish to understand such intelligence in nature, or to give such capacities to our robots, we will have to understand ontological change in individuals and in communities. Differences between precocial and altricial species are relevant, as explained below.

2.3. Assessing proposed new ontologies

Not all proposed extensions to our ontologies are equally good: Priestley's phlogiston with its negative mass lost the battle against Lavoisier's ontology permitting new processes in which oxygen in air combines with solid substances when they burn to produce solid oxides weighing more than the original solids. Some ontological victories are only temporary: Young's wave-based ontology for light demolished Newton's particle-based ontology, but the latter received a partial revival when the ontology of physics was extended to include photons.

As those examples show, it can be very difficult to decide whether a proposed new ontology is a good one. In part that is because testing is not always easy. Some extensions remain hypothetical for many years until they are explained in the framework of a broader theory: for example it took many years for the existence of genes and some of their properties to be explained using biochemistry.

The difficulty of choosing between rival theories on the basis of experiment and observation led Lakatos (1970) to develop his theory of progressive and degenerating research programmes, whose relative merits can only be distinguished in the long term. During the interim some people will use one new ontology while some prefer an alternative change, and some claim that the previous ontology was good enough.

This paper is in part about the ontology required for adequate theories concerning the capabilities of biological organisms such as birds, apes and humans, and in part about the fact that some disputes in biology, psychology and AI arise out of unacknowleged differences in ontologies used by different scientists. When a group of scientists cannot think about a class of entities, properties, relations, and processes they will not be able to perceive instances of them as instances. We call this 'ontological blindness'. It can have many causes and different sorts of cures may be required. A full account of the processes by which the ontologies used by scientists change or grow is beyond the scope of this paper. However, we illustrate the process by describing some features of the ontology required for scientific investigation of intelligent animals and robots, and an application of the ontology in developing an explanatory architecture, H-CogAff, described in Section 7.

3. Ontological blindness and its cure

If some researchers are 'ontologically blind' to certain important features of naturally occurring information-processing systems, this can restrict not only their understanding of animals, but also their ability to design new biologically-inspired machines. As implied above, this is just a special case of a general problem at the frontiers of science.

A particular variant of the sort of ontological blindness we are discussing would involve attributing too simple an ontology to an organism that is treated as an information-processor. An example would be not noticing that an organism can take account of the intentions or emotional states of conspecifics, in addition to taking account of their location and physical movements. The ability to monitor and perhaps modify one's own information processing (as opposed to one's own movements or temperature changes, for example) might also go unnoticed by observers, whether they are scientists looking for explanatory theories or robot designers looking for inspiration. Partial ontological blindness may occur when scientists notice a phenomenon (e.g., vision) but misconstrue it, using the wrong ontology to describe it, e.g., thinking of vision as merely providing information about physical shape, structure, colour, and texture (Marr, 1982), and ignoring perception of affordances (Gibson, 1986).

3.1. Some consequences and causes of ontological blindness

A consequence of not noticing the more abstract capabilities in organisms (or the need for them in robots) is using too simple an ontology in explanatory theories (or design specifications). This can sometimes either be caused by, or cause, adoption of formalisms or information-encoding mechanisms that are not capable of supporting the diversity required for the ontology. This is linked to inadequate theories about mechanisms for acquiring, storing, transforming and using that information. Thus ontological blindness can be linked to paucity of formalisms and paucity of information-processing mechanisms discussed by theorists and designers.

All of this will be familiar, or at least obvious once stated, to many biologists, psychologists, neuroscientists and roboticists. For instance it is totally consistent with the methodology reported in Arbib's WGW'02 paper (Arbib, 2002), which we read only after producing our paper for the conference and which provides many detailed examples of varieties of information processing in organisms and robots. Our objective is not merely to contribute to the sort of detailed analysis presented by Arbib, but to present a conceptual framework which can help us to characterise the aims of such research and to draw attention to gaps and unanswered questions that can usefully drive further research: i.e., discovering types of ontological blindness in order to remedy them.

3.2. Ontological blindness concerning organisms

Which specific forms of ontological blindness may be hindering progress both in understanding organisms and in designing robots? A researcher who thinks the function of visual systems is merely to provide information about lower-order physical phenomena such as geometrical shapes, motion, distances, and colours (Marr, 1982), or statistical correlations between image patterns, may never notice situations where vision provides information about abstract relationships between relationships (Evans, 1968), information about affordances, e.g., graspability, obstruction, danger, opportunity (Gibson, 1986), or information about causal relationships that produce or prevent change, e.g., a rope tied to a post and a stick constraining motion of the stick (Kohler, 1927).

Similarly, a researcher who thinks the only goals organisms can have, are to achieve or prevent certain 'low-level' physical occurrences, such as maintaining a particular temperature or hormonal concentration, approaching a physical location, may never notice other sorts of goals whose description is more abstract, such as the goal of trying to work out what caused a noise, or the goal of improving a way of thinking about certain problems, or the goal of finding out what another animal is looking at.

Someone who thinks that all learning is learning of associations may fail to notice cases where learning includes extension of an ontology, or development of a new representational formalism (Karmiloff-Smith, 1996). Chomksy's (1959) attack on Skinner provides many examples.

Whether or not these latter forms of learning can be or are *realised in* or *implemented in* a purely associative learning mechanisms is beside the point; a theorist who 'sees' only the associative mechanisms will be ontologically blind to other forms of learning, just as a scientist who 'sees' only atoms, molecules, and their interactions will be ontologically blind to muscles, nerves, digestive systems and homeostatic mechanisms in animals.

We believe that ontological blindness of types mentioned above has hampered work on biologically-inspired machines. However, ontological blindness need not be permanent: a recurring feature of the history of science is a process of extending the ontologies employed, thereby changing what becomes not only thinkable but also observable, somewhat like learning to read a foreign language in an alien culture. A language for talking about different ontologies requires a meta-ontology. We will try to show how a good meta-ontology for information-processing architectures can drive fruitful ontological advances. Our proposed first-draft meta-ontology is a generative schema.

3.3. Architecture-based exploration of ontologies

We suggest that one useful way (not the only way) in which we can overcome some kinds of (temporary) ontological blindness is to use a *generative schema* for a class of architectures defining a space of possible designs to be related to a dual space of possible niches for which such designs may be more or less 'fit' in different ways. If this provides us with a framework for systematically extending our ideas about architectures, functions and mechanisms, it may, to some extent, help to overcome ontological blindness. It may also help us generate a unified terminology for describing designs and explanatory theories.

Suppose we find that a particular explanatory architecture is inadequate to explain some capabilities, e.g., visual problem solving, or competence in certain games. We can then use the architecture-schema to generate alternative architectures that differ in terms of forms of representation, forms of reasoning, and forms of control and see if one of them comes closer to the required capabilities.

Alternatively, if we find that a particular explanatory model fails to replicate some observed behaviour, and we cannot find any change that works, this may prompt us to ask whether that is because there are aspects of the niche that we have not yet identified (e.g., forms of perception, or kinds of goals the organism needs to have, or ways of learning that we have not considered). This can lead to an extension of our ontology for niches (an extension of 'niche space') which then leads us to look at the proposed architecture and consider new ways of modifying it in ways suggested by the schema, e.g., making more functional divisions within the architecture, considering using new forms of representation or new mechanisms within existing components, or adding or removing forms of communication between components in the architecture. This, in turn, may lead us to consider a kind of architecture not previously thought of, or may lead to the idea of a new kind of function for a sub-mechanism, promoting a search for suitable sub-mechanisms.

Evaluating the suitability of the modified architecture for the supposed niche may show that it would fit better in a different niche, and that may lead to the hypothesis that we have mis-identified the niche of the organism under study, causing us to extend our ontology for types of niche.

Moreover, by noticing how new types of states and processes can arise in the proposed modified architecture we discover the usefulness of new architecture-based concepts as explained in (Sloman & Scheutz, 2001; Sloman, 2001a; Sloman & Chrisley, 2003). This parallels the history of computer science and software engineering, in which explorations of new *designs* led to the discovery of new useful *concepts* which feed back into new designs, for instance discovering the usefulness of notions like 'deadlock', 'thrashing' and varieties of 'fairness', in consequence of moving from single-threaded to multi-threaded operating systems.

It is also possible to discover that our meta-ontology, the schema for generating architectures, is too restrictive; so one of the possible forms of advance is extending or modifying the schema. Later we describe a first draft schema, CogAff (Fig. 1), for describing a wide class of information-processing architectures for animals and robots and show how it can help to reduce such ontological blindness. We also present a first-draft particular architecture, H-CogAff (Fig. 5), proposed for human-like systems, arrived at by applying this methodology. Both the schema and the architecture are the result of many years of work in this area, and have developed gradually. Both are still in need of fur-

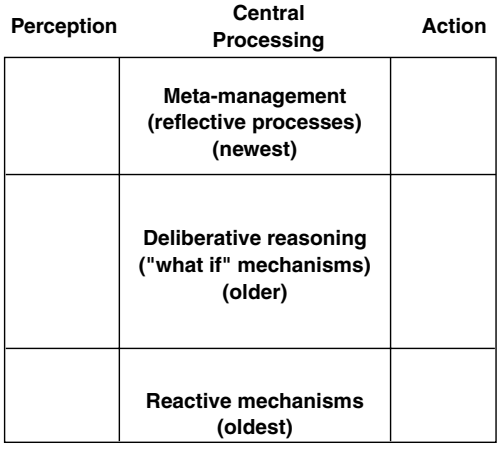

Fig. 1. CogAff schema component grid.

ther development, which will take many years of multi-disciplinary collaborative research.

4. How to avoid the problem?

One of the recurring themes in AI is that natural systems are too complex for us to design, so that an alternative is proposed: e.g., design a system that can learn and train it instead of programming it, or design an evolutionary mechanism and hope that it will produce the required result.

The ability of a new-born human infant to learn a huge variety of things it lacks at birth may at first seem to be an existence proof that the first approach works. But if we don't know what sort of learning mechanisms infants have, we may fail to design a machine with the required capabilities. An apparently helpless and almost completely incompetent human infant may in fact be born with more sophisticated architecture-building mechanisms than those available at birth in precocial species, like deer that can walk, suckle, and run with the herd within hours.

At least we know that biological evolution started from systems with no intelligence, so those who are tempted to avoid thinking about how to *design* biologically-inspired robots may instead try to *evolve* them, since animals provide an existence proof of the power of evolutionary mechanisms. But this may not lead beyond the most elementary of robots in the foreseeable future, because of both the computational power required for replicating evolution of complex animals and also the problem of designing suitable evaluation functions (Zaera, Cliff, & Bruten, 1996).

In natural evolution, implicit evaluation functions develop partly through co-evolutionary processes which alter niches. Replicating this may require simulating the evolution of many species, leading to astronomical computational requirements. Moreover, insofar as the process is partly random there is no way of knowing whether simulated evolution will produce what we are trying to replicate. Even on earth there was never any *guarantee* that penguins, piranhas or people would ever evolve. So the time required to evolve a robot like those organisms may include vast numbers of failed attempts. Perhaps explicit design, inspired by nature, will be quicker.

Moreover, from a scientist's point of view the mere *existence* of an evolved design, whether natural or artificial, does not aid our understanding if we are not able to say what that design is, e.g., what the information-processing architecture is and how it explains the observed behaviour. An engineer should also be wary of relying on systems whose capabilities are unexplained.

People working on artificial evolution have to design evaluation functions, evolutionary algorithms and the structures on which the algorithms operate. But the design of the evolutionary system does not explain how a product of such a system works. It merely provides a source of more unexplained examples, and partially explains how they were produced.

The task of trying to understand a product of natural or artificial evolution is not unlike the task of finding out how to design it, since both understanding and designing involve specifying what sorts of architecture the system has, what mechanisms it uses, what sorts of information it acquires, how it represents or encodes the information, and how it stores, manipulates, transforms and uses the information, and understanding what difference it would make if various features were changed.

5. How to attack the problem

This paper is motivated by the belief that (a) we shall have to do some explicit design work in order to build robots coming anywhere near the capabilities of humans and other mammals and (b) knowing how to design something like X is a requirement for understanding how X works. So both engineers and scientists have to think about designs.

Of course doing explicit design is consistent with leaving *some* of the details of the design to be generated by learning or adaptive processes or evolutionary computations, just as evolution in some cases pre-programs almost all the behavioural capabilities (precocial species) and in others leaves significant amounts to be acquired during

development (altricial species). In the altricial case, what is needed is higher-order design of bootstrapping mechanisms (Sloman, 2001a, 2001b). In that case, design-based explanations may produce understanding only of what is *common* to a class of individuals whose individual development and learning processes produce great diversity.

5.1. Organisms are (information-processing) machines

Machines need not be artificial: organisms are machines, in the sense of 'machine' that refers to complex functioning wholes whose parts work together to produce effects. Even a thundercloud is a machine in that sense. In contrast, each organism can be viewed simultaneously as several machines of different sorts.

Clearly organisms are machines that can reorganise matter in their environment and within themselves, e.g., when growing. Like thunderclouds, windmills and dynamos, animals are also machines that acquire, store, transform and use energy. However, unlike most of the systems studied by physical scientists or built by engineers in the past, organisms are also *information-processing* machines (some people would say 'cybernetic systems').

Many objects whose behaviour is directed by something in the environment acquire the energy from the same thing: a string pulling an object, a wall causing a change of direction, wind blowing something, etc. In each case the source of energy also determines the resulting behaviour, e.g., the direction of movement. Most (perhaps all?) organisms, however, use internal (mostly chemical) energy to power behaviour invoked by external information. Having sensed environmental features then, depending on their current state, they select useful actions, and use internal energy to achieve them, for example producing motion in the direction of a maximal chemical or temperature gradient, or motion towards a light, or a potential mate or food. (Compare the discussion of 'switching organs' in (von Neumann, 1951), for which 'the energy of the response cannot have been supplied by the original stimulus. It must originate in a different and independent source of power' (p. 426).)

5.2. Varieties of information-based control

Information acquired through sensors and the action-selection processes will be different for organisms with different niches, even in the same location, for instance a caterpillar, a sparrow, and a squirrel on the same tree branch. The use of the information will also vary with the internal state, e.g., selecting motion towards food when hungry or towards other things otherwise. For most non-living things the influence of the environment is purely through physical forces, and resulting behaviour is simply the *resultant* (vector sum) of the behaviours produced by individual forces. In contrast, an information-processing system can consider options available in a situation, and then decide not to act on some external information when there are conflicting needs.

But we need to be careful about familiar words like 'consider' and 'decide', for in the context of the simple organisms lacking human-like deliberative capabilities, *consideration* of an option may merely amount to activation of some set of neurons capable of producing the appropriate behaviour, and *deciding* may amount to no more than the result of a competitive 'winner-takes-all' process among clusters of neurons. We could call that a 'proto-deliberative' system, found in many organisms capable of producing different behaviours depending on the circumstances and capable of switching discontinuously between behaviours as the situation changes continuously, e.g., a predator approaches, as discussed in (Arbib, 2002).

In a more sophisticated organism (or robot), *considering options* may involve building structural descriptions of the options, deriving consequences of each of them, deriving consequences of the consequences, building descriptions of pros and cons, and then using some rule or algorithm to select the best option. The organism may also store the reasons for the selection, in case they are needed later if the situation changes. The reasons may also contribute to a learning process. This is an example of what we call a 'deliberative system'.

Deliberative systems come in many forms, though they all involve some ability to represent

non-existent possibilities (Sloman, 1996a), which we can summarise as *the ability to do 'what if' reasoning*. They can differ in the variety of types of non-actual possibilities that they can consider and select, the variety of forms of representation that they can use for this purpose and the variety of uses to which they put this capability, e.g., planning future actions, explaining observed events, predicting what another agent will do, or forming hypotheses about unobserved portions of the environment.

Simple versions may be able to do only one-step look-ahead and may use fixed formats for all the possibilities they consider. More sophisticated deliberative mechanisms may be able to do more complex searches and use structural descriptions of varying complexity, depending on the task, and using compositional semantics as a source of generality. They may also be able to use representations of hypothetical situations to speculate about the past, about remote or hidden objects, or about unobserved explanations for observed phenomena. So deliberative processes in our sense of the phrase, are not restricted to planning and action-selection tasks.

The extra generality and flexibility required to support complex and varied deliberative processes may incur a heavy cost in brain mechanisms and prior learning of re-usable generalisations. The cost of the brain mechanisms for storing large numbers of rapidly retrieved, re-usable generalisations, and mechanisms for supporting the construction and use of temporary descriptions of many kinds may be a partial explanation of the rarity of sophisticated deliberative capabilities in animals: very few animals can exist near the peak of a food pyramid.

According to Arbib's description of a frog (Arbib, 2002), it has proto-deliberative capabilities, in our sense, though he uses the label 'deliberative'. However some of his more complex examples come closer to what we call deliberative architectures. The choice of labels is unimportant. What is important is to understand the architectural differences and their implications. We still have much to learn about the space of design options and their trade-offs.

5.3. Information-processing architectures

Investigating these phenomena in order to design robots that replicate them, requires deep theories regarding various types of internal processing. Obviously biological evolution produces many changes in physical design. Not so obviously there are changes in information-processing capabilities, which are sometimes far more dramatic than the physical changes. For example, apes and humans are physically very similar (i.e., there are simple structural mappings between most of their physical parts) whereas some of their information-processing capabilities are very different as shown by their behaviour and its products (their extended phenotype). On a small scale their movements may be similar: walking, climbing, jumping, grasping, eating, etc. But on a large scale there are huge differences insofar as only humans, given a suitable environment, make excavators, cranes, skyscrapers, aeroplanes, farm many kinds of food, do mathematics and write poetry. Creatures with structurally similar bodies can have structurally very dissimilar minds. [6] Furthermore, given that brains are highly complex and therefore extremely sensitive to boundary conditions, even organisms with identical *brain structure* can have very different minds. A level of characterisation above the physical, anatomical level will do better at modelling this substantial difference by representing it with a substantial difference in the characterisation itself.

5.4. Hidden differences in information-processing

Given its abstract, non-physical nature, information-processing may be difficult to detect in natural systems using observational techniques usual in the physical sciences.

Even when similar behaviours are observed in different organisms it does not follow that the behaviours are the outcome of similar internal processes (Hauser, 2001). Less obviously, similar

[6] Some may argue that the minds have similar architectures, but differ only in their information *content*. However that does not explain why the same content cannot be acquired by both sorts if their minds have similar architectures initially.

behaviours in the same organism at different stages of development, or training, e.g., grasping, breathing, smiling, visual-tracking, may be products of very different internal processes.

Furthermore, as argued in (Sloman, 2001b), two organisms in the same environment may perceive radically different things. For example, a deer and a lion apparently gazing at the same scene will not necessarily see the same things, since their niches and affordances differ substantially. In particular, altricial species (which are born under-developed and almost helpless, e.g., lions) may develop major aspects of their visual capabilities during early development whereas adults of precocial species (born with a more advanced collection of capabilities, e.g., deer, sheep) have simpler capabilities mostly produced by their genes – e.g., enabling new-born grazing animals to stand, walk, find and suck nipples or even run with the herd within hours of being born. Hunters, nest-builders and berry-pickers appear to perform intricate actions taking account of multiple constraints and affordances, whereas the actions of grazers are not dependent on understanding such complex structures and processes in the environment. This could explain why the kind of genetic encoding of affordance detection that suffices for grazers is inadequate for altricial species.

5.5. Varieties of information-processing systems

We still have much to learn about information-processing systems. The simplest kinds can be described in terms of homeostatic feedback loops or hierarchical control loops, possibly characterised by sets of partial differential equations. But we also know that there are many information-processing machines (including parsers, planners, problem-solvers, operating systems, compilers, email networks, theorem provers, market trading systems, chess computers) whose most useful explanatory description does not take that form.

There is no reason to assume that all biological information processors will turn out to be simply large collections of analog feedback loops, even adaptive ones: work in AI in the last half century demonstrated that there can be much more powerful alternative forms of information-processing and control, that are particularly useful for some tasks, for instance those in which it is not immediately evident what the consequences of each available action are – as in most tasks where a complex structure subject to many constraints has to be built from diverse components. But we do not yet have a good overview of all the alternative mechanisms, or their strengths and weaknesses, and that makes theory construction very difficult.

6. How to describe information processors: niches and designs

It is only recently that scientists and engineers have begun to understand requirements for investigating and modelling information-processing systems. Using an overly restricted conceptual framework can constrain the questions asked and the theories proposed in the study of humans and other animals. This can also lead to a narrow view of robot functionality (Braitenberg, 1984).

It is also common to use a restricted notion of computation, defined in terms of something like a Turing machine (Sloman, 2002b). An alternative is to treat 'computation' and 'information-processing' as very broad terms applicable to a wide range of types of control systems. For instance, we do not exclude information-processing systems that contain continuously varying states, whereas Turing machines and their equivalents must be discrete. [7]

6.1. Towards an ontology for agent architectures: CogAff

We are attempting to complement and broaden existing approaches by developing a schematic framework called 'CogAff', [8] depicted in Fig. 1, for comparing and contrasting a wide range of information-processing architectures (typically virtual machine architectures, which will not necessarily map in any simple way to the underlying

[7] See also talks 4 and 22 in http://www.cs.bham.ac.uk/research/cogaff/talks/.

[8] Described in previous papers (e.g., Sloman, 2000b, 2001a, 2002a).

physical architecture). Although investigation of specific architectures is important, scientists and engineers also need to understand the class of possible designs from which they make selections, lest they unwittingly ignore important alternatives, and reach mistaken conclusions. In order to understand the trade-offs between alternatives we also need a generative framework for talking about types of niches, or sets of requirements, relative to which architectures can be evaluated, possibly using multiple dimensions of evaluation, as noted in (Sloman, 2000a).

6.2. Terminological inconsistencies.

The CogAff framework permits combinations of mechanisms producing concurrent processes roughly classified as reactive, deliberative and meta-management (sometimes labelled 'reflective') processes. Unfortunately there is much terminological confusion among researchers studying architectures. Some people use 'reactive' to exclude state changes. We don't. Some distinguish reflexes from reactive mechanisms, whereas we treat them as a subset of reactive mechanisms. Our use of 'reactive' excludes only deliberative processes involving explicit consideration and comparison of possible alternatives of varying complexity, whereas proto-deliberative systems, described earlier, are classified as reactive. (Perhaps it would be better to use some intermediate categories.) A reactive system may also be able to invoke and execute stored plans, where the plans have been produced by evolution, by training, or by another part of the system. Compare (Nilsson, 1994). In contrast, some people describe anything that makes choices as 'deliberative'.

There is no question of trying to prove that our terminology is right or wrong. The important thing is to understand the variety of types of mechanisms that are available and the different ways in which they can be combined in an integrated architecture. We offer the CogAff schema only as a first draft, very sketchy, starting point, illustrating the more general point that we need a generative schema.

Not all three-layered architectures described in the literature are the same, even if the diagrams look the same and similar-sounding terminology is used. For instance, an architectural layer labelled as 'deliberative' is often regarded simply as a planning system, whereas our notion of a deliberative mechanism includes the ability to consider alternative explanations of some observed facts or to speculate about distant or hidden objects. Some people use 'reflective' to refer to an architectural layer containing mechanisms that observe what happens when plans are executed in the environment, and perhaps learn from the results, whereas we treat that as a feature of what we call the 'deliberative' layer. The processes in our third layer, the meta-management layer, described in (Beaudoin, 1994), are concerned with observing, evaluating, and controlling *information-processing* processes within the rest of the architecture. Insofar as this requires manipulating information about information, we describe it as a *meta-semantic* function. The representational requirements for meta-semantic competence go beyond the requirements for representing physical states and processes within the agent or in the environment, e.g., because of the need to support referential opacity: expressions that fail to refer can be part of a meta-semantic description. For instance, I can think the person following me wants to mug me, when there is no person following me. Moreover, I can later describe myself as having had the mistaken thought.

Some researchers would restrict meta-management, or reflection, to self-observation of processes within the central cognitive system, whereas our notion of meta-management includes the ability to attend to intermediate structures and processes in perceptual and action mechanisms, some of which may have semantic content. For instance you can attend to an aspect of your visual experience in which one object obscures a part of another object that you are trying to see. Similarly you can attend to whether a movement you are making is done in a relaxed or tense way, or with or without precise control. (Relaxed precision is a requirement for many sporting and artistic achievements.)

In (Sloman & Chrisley, 2003), we have tried to show how this ability to monitor internal information-processing states can involve mechanisms that

account for many of the features of what philosophers refer to as 'qualia'.

There may be both reactive and deliberative meta-management processes, since what distinguishes them is what they are concerned with and not what mechanisms they use. [9]

In the case of animals, including humans, these processes use mechanisms and forms of representation that evolved at different times. Within this framework we can analyse the trade-offs that might have led to evolution of hybrid systems with various subsets of the components accommodated in the CogAff schema. However, we stress that our three-way distinction between different architectural layers is a first crude sub-division, and detailed investigations of evolutionary and developmental trajectories are likely to show interesting intermediate cases, requiring a more refined ontology.

A particularly interesting possibility suggested by this framework is that the ontology and forms of representation used for perceiving and reasoning about information processing in *others* may have co-evolved with and overlapped with those used for *self* monitoring and control, i.e., meta-management, though there are many who believe that social uses of meta-semantic competence must have preceded self-directed meta-management, or self-consciousness. (The 'simulative reasoning' approach to belief and plan ascription, favoured by some AI researchers is consistent with both views.)

6.3. Layers in perceptual and action systems: multi-window perception and action

The three-way distinction does not apply solely to central processing, but allows us to distinguish perceptual and action sub-systems according to whether or not they have components that operate concurrently at different levels of abstraction related to the three architectural layers. In Sections 1 and 2 we pointed out that scientists can view the same subject matter on different levels of abstraction. This ability is not restricted to scientists, nor even to humans. We all perceive on the meta-mental level when we see the state of mind of another person (seeing someone as happy, angry, in pain or attentive). There may be cases of non-human organisms perceiving on the deliberative or meta-management levels, as opposed to being capable only of doing feature detection or pattern recognition of the lowest order. In (Sloman, 2001b) and (Sloman & Chrisley, 2003) we labelled the two options 'multi-window' and 'peephole' perception. The same contrast can apply to action systems. The possibility of layered perception and action systems should be reflected in any attempt to characterise the space of possible architectures for biological or robot intelligence. Later, in Section 7.2, we discuss an objection to this idea.

6.4. The CogAff grid: a first draft schema

Fig. 1 schematically indicates possible types of concurrently active sub-mechanisms within an architecture. Information can, in principle, flow in any direction between boxes, or between sub-mechanisms within a box. Thus data-driven perception of high level features involves information flowing up the left hand box, undergoing different kinds of processing to meet the needs of different layers in the central box. In contrast, top-down processing could involve information flowing down, because more abstract percepts, and prior information in different central layers, can influence processing of low level features. Simple reflex actions could involve information flowing directly from the low level perceptual layer to the low level action layer. More sophisticated reflexes could involve high level, abstract, perceptual information triggering low level internal mechanisms, as happens in some emotional reactions, for instance. Proprioceptive information would come from some point in the action hierarchy to a central mechanism, and so on.

Not all architectures include mechanisms corresponding to all parts of the grid. Different architectures will have different components and different communication links between components. For instance, some may have only the reactive layer (which may include several different sub-layers,

[9] To add to the confusion, everything has to be ultimately implemented in reactive processes, otherwise nothing would ever happen.

as in most subsumption architectures, indicated in Fig. 2). Some may include 'diagonal' information links, for instance high level perceptual processes triggering low level internal reactions (which may be part of what happens in some aesthetic experiences). Additional mechanisms and information stores that do not fit neatly into the CogAff boxes may be needed to support the mechanisms in the boxes.

6.5. Omega architectures

A popular sub-category of the CogAff schema is what we call an *Omega architecture*, depicted in Fig. 3, which uses only a subset of the possible sub-mechanisms and routes permitted by the schema, forming roughly the shape of an Omega: **Ω**. Omega architectures use an information pipeline, with 'peephole' perception and action, as opposed to 'multi-window' perception and action described in Section 6.3. The 'upward' portion of the pipeline generates possible actions triggered by the sensory input. Selections among options are made at the top and the chosen options are decomposed into low level motor signals on the 'downward' pathway. The 'contention scheduling' architecture of Cooper and Shallice (2000) has this sort of structure, as does the three-layered architecture of Albus (1981) which superficially resembles the H-CogAff architecture described below, but turns out on closer examination to be an Omega-type architecture with something called 'the will' at the top selecting among options generated at lower levels. People who have not understood the requirement for concurrent hierarchical processing within perceptual and action sub-systems (what we called 'multi-window' perception and action) tend to take the Omega structure for granted, though they may propose different sorts of intermediate mechanisms generating options and different sorts of 'top-level' decision-making mechanisms.

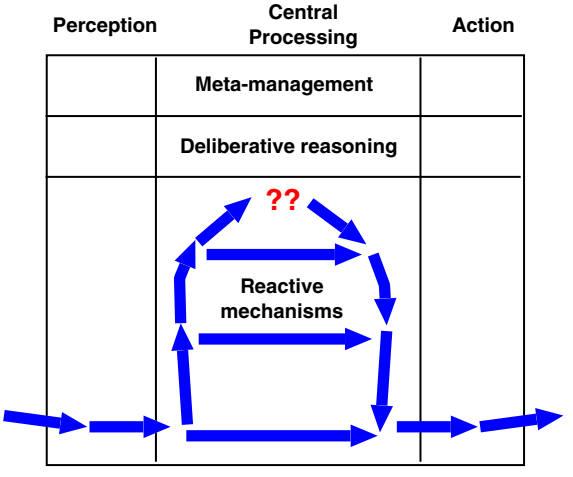

Fig. 2. Common subsumption architectures are subsumed by CogAff.

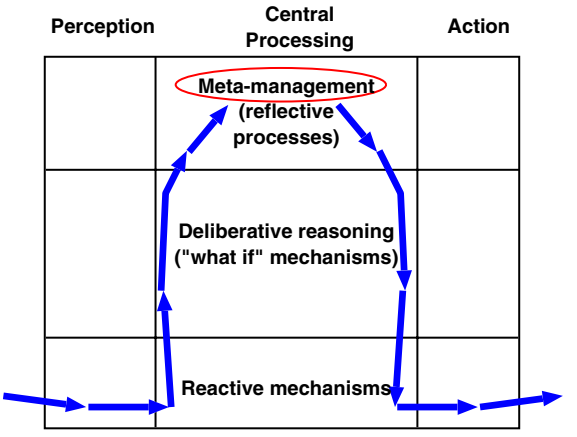

Fig. 3. The 'Omega' type of architecture uses a pattern of information flow between layers in the CogAff schema reminiscent of a Greek letter Ω.

6.6. Alarm mechanisms

Some architectures include one or more 'alarm mechanisms' (Fig. 4), i.e., reactive sub-systems with inputs from many parts of the rest of the system and outputs to many parts, capable of triggering global reorganisation of activities, a feature of many emotional processes. Alarm mechanisms may be separate sub-systems running in parallel with the systems they monitor and modulate, or they may be distributed implicitly within the detailed sub-mechanisms, e.g., in conventional programs using very large numbers of tests scattered throughout the code. The former, more modular, type of alarm sub-system may allow more global forms of adaptation and more global kinds of control options when dealing with emergencies, at the cost of architectural complexity.

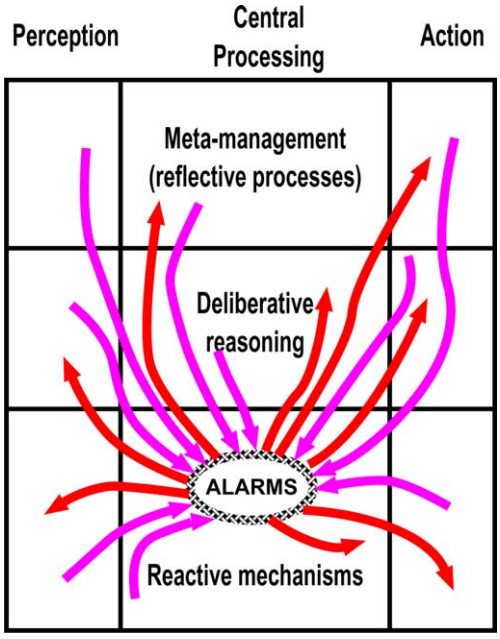

Fig. 4. Grid with 'alarm' mechanisms.

6.7. An objection considered

An objector might ask: how can one distinguish architectures that have input and output only at the lowest level (like the 'omega' architectures discussed in Section 6.5) from those with input and output on multiple levels, given that all high-level input and output must be realised by low-level input or output? Surely, when an organism receives the high-level visual input that there is food nearby, it does so by virtue of receiving low-level input, e.g., photons hitting retinal cells and producing image features such as intensity, colour, texture, optical flow, etc., and variations therein. Similarly, executing a high-level plan to return to one's mate, by following a route, requires executing a sequence of low-level behaviours and muscle movements. This line of thought suggests that something like the 'Omega' model (Fig. 3) is the only possible architecture for organisms that perceive and act on higher levels. In such architectures (a) all input received is low-level, although possibly trans-

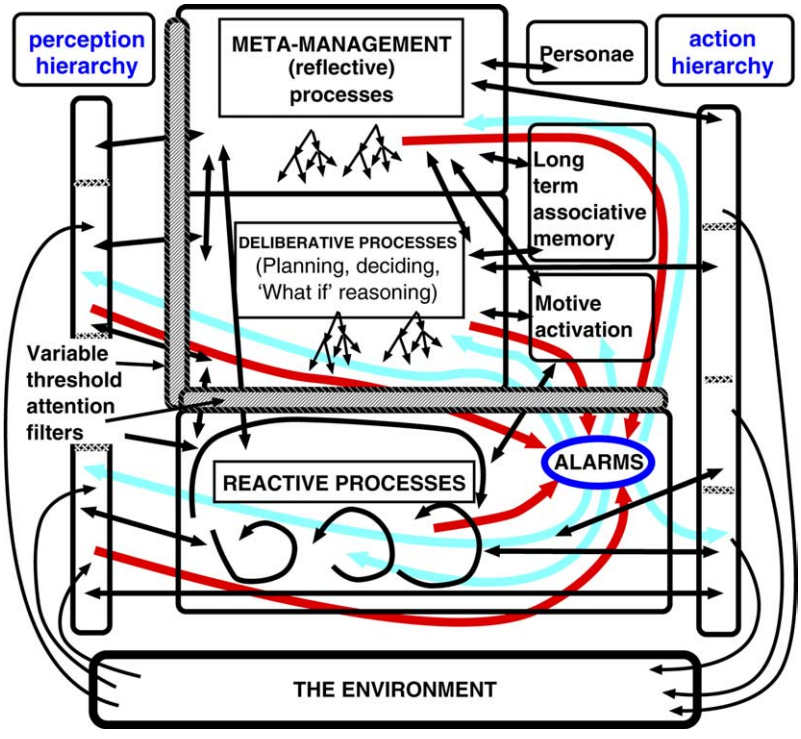

Fig. 5. The H-CogAff Architecture.

formed into higher-level categories during deliberation, etc., and (b) all output is low-level, although possibly the result of deliberation involving more abstract characterizations of action,

This argument ignores good reasons for distinguishing between the Omega architecture and architectures involving true, multi-level perception and action (such as H-CogAff, Fig. 5). The latter satisfy specific requirements on the high-level perceptual processes (Sloman, 1989). For example, for multi-level perception, we would require there to be higher-level representations (such as affordances involving more abstract ontological categories) which are the product of dedicated perceptual modules that:

(a) have the *function* of producing said representations (e.g., they evolved, or were designed, to do this, and this is all they do, unlike general-purpose inference mechanisms);
(b) run in parallel with other processes and partly independently of general-purpose central reasoning, learning, planning mechanisms; and
(c) use some special-purpose, modality-specific, forms of representation, e.g., higher-level representations that are in registration with low level sensory arrays that are different for different sensory modalities, vision, hearing, touch, etc. (Compare the 'place-tokens' in (Marr, 1982).)

And similarly, *mutatis mutandis*, for multi-level action. Note that the modularity assumed here is weaker than, e.g., Fodor (1983) in that the modules need not be cognitively impenetrable nor totally encapsulated. That is, high level and low level visual processes can be very much influenced by central processes, including current goals and problem contexts, and still be modular and therefore distinct from an Omega architecture.

It is not uncommon for AI visual systems to have dedicated mechanisms for extracting some higher level information from low level visual data, for instance, classification and location of 2-D regions, or 3-D objects, or 4-D trajectories of 3-D objects, or parsing in the case of reading sentences. In the case of H-CogAff we postulate more subtle and sophisticated visual processing, for instance categorising other agents in terms that use meta-semantic ontologies, e.g., seeing another as happy, or sad, or as intending to do something, as explained in (Sloman, 1989). We have found no mention of this sort of thing in connection with Clarion (discussed below in Section 8) or any other well-known AI architecture, although the growing awareness of the importance of perceived affordances, following Gibson (1986) points in this direction.

Another way to distinguish Omega-style from true multi-level perception and action would be to require input and output mechanisms to be non-deliberative. On this view (which is probably inconsistent with the module-based approach just described), if deliberative mechanisms are involved in the transformation from low-level to high-level input, and from high-level to low-level action, then the Omega architecture best describes that organism. If, however, the low-level input of an organism is transformed into high-level categories by way of non-deliberative, automatic, blind, reactive processes, that are incapable of considering and comparing alternative high-level interpretations of the same data, then that organism can be said to be engaging in true, multi-level perception.

An intermediate case would be dedicated perceptual mechanisms which, like parsers, sometimes need to *search* for coherent global interpretations of collections of locally ambiguous sensory data. This may have some similarities with the cognitive processes involved in searching for a plan, a proof or an explanation. But if the required functionality is implemented in mechanisms that are dedicated to processing of sensory input in order to produce higher level percepts, that is consistent with the label 'multi-level' perception, in contrast with an Omega architecture.

We conjecture that a great deal of human perceptual learning involves developing such dedicated perceptual and action mechanisms, e.g., in learning to read, learning to understand spoken language, and learning many athletic skills.

Similar remarks can be made about multi-level action mechanisms. But note that these architectural features are independent: An architecture may have multi-level perception without having multi-level actions, and, like Clarion (discussed

in Section 8) may have multi-level output without having multi-level perception.

All of the kinds of architectures we have been discussing are 'virtual machine' architectures as explained in Sections 1.1 and 2. This implies that there need not be any simple mapping between the components of the architectures and physical components of brains or computing machines in which the architectures are implemented (or realised). This means that empirical investigations testing claims about architectures used in animals will be very dependent on indirect evidence.

7. Links to empirical research

Using the framework developed in previous sections, including the notion of a virtual machine architecture and the notion of a generative schema for a class of architectures, of which CogAff is a simple example, we can study organisms by trying to identify architectures that contain components and information linkages able both to explain observed capabilities and also to suggest research questions that will extend what we know about the organisms, generating new requirements for explanatory architectures.

7.1. CogAff and emotions

For example, we have shown in (Sloman, 2001a) how a full three level architecture, of the sort represented within the CogAff schema, can explain at least three different classes of emotions found in humans namely *primary* emotions involving alarm mechanisms in the reactive layer, *secondary* emotions involving reactive and deliberative layers triggering alarms which modulate processing in both layers, and *tertiary* emotions in which alarm mechanisms and other mechanisms disrupt the meta-management layer, leading to loss of control of attention.

More detailed analysis based on this schema can lead to a richer, more fine-grained classification of types of emotions and other affective states, including desires, preferences, likes, dislikes, attitudes, and moods. Different types of emotions, all depending on the ability of one part of the system to detect a need to interrupt, re-direct or modulate another part, can be distinguished by distinguishing different sources of alarm triggers, and different components in which the alarms can cause disruption of some kind, as well as different time-scales of operation, and whether there are secondary effects, such as the meta-management system being disturbed by noticing a disturbance in another part of the system, or even in itself as described in the case of human anger in (Sloman, 1982). These processes can also be related to mechanisms that activate and maintain or deactivate motivations and moods.

It is worth noting that emotions as we construe them do not require a special 'emotion mechanism' within the architecture, as proposed by many researchers. Rather the three types of emotions occur as the result of the operation of and interactions between mechanisms whose primary functions are not best described as being 'to produce emotions'.

Organisms with only a subset of the architectural layers will not be capable of having the variety of emotions and other states that are possible according to the CogAff schema. Obviously if insects lack a deliberative layer they will not be able to have emotions (such as regret!) that require 'what if' representational capabilities, as most humans can. If human infants lack deliberative mechanisms they too will be unable to have mental states that depend on them. Various kinds of disorders may also be related to different parts of the architecture. Barkley (1997) discusses meta-management architectural features relevant to disorders of attention, though without using our terminology.

The generic CogAff framework allows many variations in conforming architectures, including both simpler, insect-like architectures, and more complex additional mechanisms required for more sophisticated deliberative and meta-management processes. In (Sloman, 2001a) and other papers listed in the references, we outline such an elaborated instance, the H-CogAff architecture, illustrated sketchily in Fig. 5. There is much to be said about the additional components required

for all of this to work, but space constraints rule that out here. [10]

7.2. CogAff and vision

Another application of these ideas concerns theories of perceptual processing, including vision. For instance, if these ideas are correct, then (Marr's, 1982) specification of the functions of vision where he describes the 'quintessential fact of human vision – that it tells about shape and space and spatial arrangement', leaves out important types of visual processing, including the perception of various kinds of affordances, as argued in (Gibson, 1986, 1989).

7.3. CogAff layers and evolution

Although the layers and columns of the CogAff schema need not correspond to anatomically distinct components of an organism, it is consistent with such differentiation. Furthermore, the fact that the layers in a particular organism evolved at different times might make such differentiation likely. It follows that if, as we conjecture, sensory inputs in humans and some other animals are processed concurrently at different levels of abstraction, with information from the different levels transmitted concurrently to different parts of the architecture, which use the information for different tasks, then we can easily explain empirical results that have led some scientists to postulate different perceptual pathways (e.g., Goodale & Milner, 1992), though we would predict more diverse pathways than empirical evidence suggests so far. Likewise if the ability to be aware of and to report visual processing depends on the meta-management layer getting information about intermediate structures in the visual system, then we easily explain the possibility of blindsight (Weiskrantz, 1997) in a system where some connections to meta-management are damaged while some visual processes remain intact for instance in reactive mechanisms.

By analysing possible changes within the different levels and different links between the levels, we can identify many different possible kinds of adaptation, learning and development, inspiring both new empirical research and new kinds of designs for self-modifying architectures.

8. Case-study: applying CogAff to Clarion

In Section 6.4, we explained how the architectural ideas of Albus, Brooks, Shallice and others can be located within the CogAff schema, at least as regards their high level structure. Here it may be instructive to apply the schema to yet another cognitive architecture, Clarion (Sun, 2002). While the sophistication of Clarion prevents us from doing full justice to it here, attempting to relate some of its major features to CogAff is instructive, at the very least in making clear where CogAff should be extended or modified.

The main feature of Clarion is that it is specified as operating on two levels, one reactive, mainly using subsymbolic neural mechanisms, and one using explicit symbolic mechanisms. Each level is further divided into two kinds of functionality, namely 'action-centred', i.e., procedural, vs 'non-action-centred', i.e., declarative. There are further functional subdivisions between short term and long term stores, goal stacks and other mechanisms. Although such distinctions are not part of the top level CogAff framework, it is to be expected that many instances of CogAff would include such distinctions, as the H-CogAff architecture does. In particular, insofar as the deliberative layer is defined in terms of the ability to construct structurally varied descriptions of hypothetical processes or situations, in formulating plans, predictions, explanations or hypotheses, it must include both generic long-term information and also a re-usable short term workspace in which the long-term memory is applied to the current problem. So at first sight Clarion is a special

[10] At present there is no complete implementation of H-CogAff and not even a complete specification. However partial implementations of aspects of the architecture can be found in PhD theses by Luc Beaudoin, Ian Wright, Steve Allen, Catriona Kennedy, and Nick Hawes, available at http://www.cs.bham.ac.uk/research/cogaff/phd-theses.html. There is also work in progress by Dean Petters using aspects of H-CogAff in a model of aspects of attachment in infants.

case of CogAff. It is in some ways a simpler instance than H-CogAff because it excludes some of the features of H-CogAff (including multi-window perception and action), and in some ways more sophisticated because it specifies elaborate learning mechanisms.

One important difference is that Clarion's bottom level is restricted to neural mechanisms (using back-propagation) whereas the CogAff schema allows both neural and non-neural reactive mechanisms, for instance implemented as forward-chaining condition-action rules, like Nilsson's 'teleo-reactive' systems (Nilsson, 1994). As far as H-CogAff is concerned, we leave open whether there are many kinds of reactive mechanisms, including reactive rule-sets, or whether they are restricted to neural, distributed representations. (There is no evidence that animal brains use back-propagation, a major feature of Clarion's neural mechanisms.)

8.1. Multi-level perception and action in H-Cogaff and Clarion

Input into Clarion is conceptualised as ordered pairs of dimensions and values, which corresponds to something like arrays of sensor values. While this may formally allow for multi-layer perception (Section 6.3), it does not guarantee it.

Clarion explicitly allows for both primitive actions and complex structures composed of such primitives in the top layer. Moreover, some parts of Clarion such as the motivational sub-system may have both high level structured input and low level sensory input. Furthermore, although most cognitive architectures include processes that modify their own working memory and goal structure, Clarion goes further and explicitly conceptualizes these *as* actions (albeit 'internal' ones, as distinct from normal, 'external' actions). Clarion is capable of multi-level actions, but does not, as far as we can tell, include multi-level ('multi-window') perception. To that extent it is like an Omega architecture on the input side, but not on the output side.

An architecture as sophisticated as Clarion may require a separate CogAff-style analysis of each of the major components of the architecture. For example, Clarion's action-centred subsystem might have low-level input and multi-level output, but other subsystems of Clarion may be different.

8.2. Meta-cognition and meta-semantics

The discussion of such matters raises another point: Current diagrams of the CogAff schema can be misleading when talking about 'higher' or 'lower' levels of representation or processing. The single vertical dimension can indicate, depending on context, one of three different notions of 'higher' vs 'lower'. The vertical shift between layers one and two indicate a move from processes that are merely reactive to processes that are (reactive but also) deliberative. The vertical shift from layer two to layer three indicates a different shift, from non-meta to meta-management processes. The vertical dimension in the parts of the diagram depicting input and output may indicate either of the previous two distinctions, or a more general notion of 'abstractness', as was assumed in our earlier discussion of multi-level perception and action. Finally, the vertical dimension is sometimes taken to indicate phylogenetic order, with layers on the bottom being older. Of course these different interpretations of the vertical dimension are related, since, for example, in nature reactive mechanisms evolved first and are found in all organisms, whereas meta-semantic mechanisms seem relatively new and relatively rare.

Concerning meta-management, it can be difficult to locate architectures such as Clarion with respect to this distinction. There is a feature of meta-management that we believe often gets lost in discussions of reflection (e.g., Norman, Ortony, & Russell, 2003 Minsky, Singh, & Sloman, 2004), namely *meta-semanticity*, introduced in Section 1.2 The representations employed by reactive and deliberative mechanisms have semantic content, but the semantic content is typically about objects, relationships and processes in the physical environment, or in the body. The evolution of meta-management required the development of forms of representation that refer to things that have semantic content *as such*, that refer to those contents themselves, and that refer to the processes and relationships involving said semantic entities. For example, describing a planning process, or

the current state of a perceptual system, or construing some physical behaviour as execution of a plan, all require meta-semantic capability. (Compare our earlier discussion of second- and third-order ontologies in Section 1.2.)

It seems to us that much published work which makes reference to a 'reflective' architectural layer does not do justice to the notion of meta-semantic capability either because the importance of such a capability for certain kinds of reflection is not appreciated, or the required increase in representational power required for such is underestimated. For example, a distinction can be made between (merely) extensional and (both extensional and) intensional self-reference. In case of purely extensional self-reference, one uses the same referential means one uses for referring to things other than oneself, but in a way which happens to achieve self-reference. This occurs when Joe Bloggs uses the name 'Joe Bloggs', or when a process like 'Top' in Unix, which happens to have ID 23 say, displays the properties of all running processes, including those of process 23. [11] Extensional self-reference is easily achieved in artificial systems, but, we maintain, is inadequate for modelling sophisticated forms of meta-management. For that, intensional self-reference is required: Representing oneself *as* a representational thing, which in turn requires the application of (some subset of) the concepts of semantics, truth, satisfaction, meaning, etc. [12] As far as we can tell, meta-cognition in Clarion (setting goals, focus of attention, choosing learning rules, etc.) only involves extensional self-reference; however, adding intensional self-reference would not be incompatible with anything currently in the architecture.

[11] Even these examples, however, are not quite right, as they at least involve implicit notions of what a person or a process is. Extensionally self-referential representations typically succeed in self-reference without employing any conception that they are representational. So a better analogy would be a Joe Bloggs referring to 'thing number 286', where Joe Bloggs happens to be thing number 286.

[12] We do not claim to be the first to make this observation! McCarthy, for example, has been aware of this problem for many years (e.g., McCarthy, 1979), partly because he realised very early on that meta-semantic competence is hard to accommodate within first-order logic.

8.3. Motivation and affect

The very name of the CogAff schema makes it clear that accommodating architectural differences concerning motivation, emotion and other varieties of affect is one of its principal intended uses.

While CogAff in many respects attempts to be a neutral framework within which to compare different models and theories, it nevertheless builds in some assumptions which we take to be conceptually, rather than empirically, necessary. For example, we take it that a system is in one motivational or affective state rather than another primarily because of the role that state plays in mediating between the way the organism takes the world to be and the organism's actions. Thus, an affective or motivational state is a holistic property of a system, not localizable to the state of a 'motivational module' or 'affective subsystem'. This is why we did not include an 'emotion box' in either the CogAff schema or the H-CogAff architecture: The aspects of an organism which are responsible for it being in one affective state (e.g., a particular mood or emotional state) rather than another are not, in general, distinct from the total state of the reactive, deliberative and meta-management systems, their control structures, their interactions, etc. In that sense many affective states are 'emergent' properties of interactions between mechanisms. However, that does not preclude particular affective states (e.g., having a goal, or desire, or preference) being based on some explicit structure, which may have been produced by a mechanism whose function is to produce such structures, like the 'motivator generators' in (Beaudoin & Sloman, 1993).

That said, what is to be made of architectures like Clarion, which do include a motivational module? Nothing in what we have just said prohibits, a priori, the existence of an organism which has such a module, in the sense that the organism is built so that when, say, the GET-FOOD drive is activated in the module, the necessary systemic changes which constitute being in the affective state of desiring food are thereby brought about. Indeed we would expect such motivational mechanisms to be required both in organisms and in robots. But it is important to make some observations:

- Such an arrangement is not the general case, but is a highly constrained architectural configuration (roughly equivalent to Fodor's Representational Theory of Mind) which no one at present really knows how to implement, if it can indeed be implemented. However, Clarion does not in general presuppose any such theory: it allows for important sub-states to be distributed.
- Even if localized manifestation of, say, some motivational states (e.g., explicit drives) is possible, such localization might not be possible for affective, emotional or motivational states in general.
- Independently of whether or not all affective state types can be possibly localized in such a way, it seems likely that there will always be a 'surround' of non-localized affective states which complement or even enable the motivational module.
- The module discussed thus far, which involves the *manifestation* of affective states, should be distinguished from a superficially similar kind of module which *represents* affective states to the organism, enabling it to reason about such states. For example, an organism might have the capacity to predict that if it desires something and is prevented from getting that thing, it will be unhappy, perhaps violent. If it also desired not to be violent toward its conspecifics, it might isolate itself if it thought a denial of a particular desire was imminent. Thus, explicit representation of one's motivational and emotional states, and even such things as moods, can be useful even if deliberation concerning these states does not allow one directly to control them. [13] Of course, that such modules are conceptually distinct does not prevent the possibility, at least in theory, that these modules might be realized in the same hardware. In fact, it would seem that the primary reason for an organism to evolve the former kind of affective representation would be for it to serve as a means of controlling its affective states in the light of deliberation.

[13] For some evolutionary experiments showing the advantage of such explicit states in very simple organisms see (Scheutz, 2001, 2002).

Finally, we take it to be an important feature of Clarion that it is not restricted to using a goal stack (last in first out). Goal stacks can be terribly inflexible, and is in part what led the CogAff group (especially Luc Beaudoin) to introduce instead the more general notion of meta-management. Like Clarion, Cogaff requires no commitment to goal stacks, though they could be used where appropriate. Developed taxonomies of the deliberative and meta-management layers should include this distinction.

8.4. Varieties of memory, representation and learning

The mention of Clarion's working memory, above, raises the fairly obvious point that the CogAff schema as it stands does not capture all of the information processing features in which one might be interested, such as distinctions between working, episodic and semantic memory (Clarion includes all three) and how they are related to the other aspects of the architecture.

Central to Clarion is the inclusion of both explicit and implicit, as well as both procedural and declarative, representations – the CogAff schema as it stands does not distinguish between such representational types. One might be tempted to equate procedural (or implicit) reasoning with the CogAff reactive layer, and declarative (or explicit) reasoning with the CogAff deliberative layer, but this would be a mistake. For example, explicit representations may be used in purely reactive processing (although full-fledged deliberation requires explicit representation).

Another aspect of Clarion which requires an extension of the CogAff schema in order to be accommodated explicitly, involves the variety and positioning of learning algorithms (Q-learning, backpropagation, etc.) in the architecture. Furthermore, in order to allow for the existence of an architecture which is capable of doing what we are attempting to do (and trying to get others to do) – overcome ontological blindness – any taxonomy of learning methods for cognitive architectures should include ontology revision and extension, as discussed in Section 2. In some cases

this will require the development of new forms of representation and new mechanisms for manipulating them, as in learning mathematics, music, programming, and the formal sciences. The extent to which such processes would be manifested in, or distinct from, other learning mechanisms is as yet unknown. We suspect that investigating all the kinds of learning that can occur within the CogAff sub-categories, and all the kinds of changes of causal relationships that can occur in such an architecture, will lead to new, much richer, taxonomies of types of learning and development.

While we not only accept but strongly endorse the point that much detail needs to be added to the schema, nevertheless the coarse distinctions that the CogAff schema already provides form a good place to start in taxonomizing the space of biologically-inspired information processing architectures – until someone provides a better framework.

9. Exploring design space

As we gain a deeper understanding of the space of possible architectures (design space) we can raise an ever-growing collection of questions for empirical research, and also more sensibly select designs for specific sorts of biologically-inspired robots. This task is enriched by relating it to the study of different types of niches, and the relationships between designs and niches of different sorts, helping us to understand how biological evolution works, and possibly helping us design better artificial evolution mechanisms (Zaera et al., 1996).

Many people have pointed out discrepancies between extravagant claims for AI in the 1960s to 1980s and the actual achievements. There are several different sorts of explanations, including excruciatingly slow CPUs and tiny memories. More interestingly, ontological blindness of researchers in AI and Cognitive science led to over-simplified views of the problems to be solved; for instance, assuming that biological visual systems have the sole or main function of producing geometrical descriptions of the environment, and failing to notice that the variety of concurrent interacting processes involved in natural intelligence, rules out the simple sense-think-act cycles of many AI systems.

In particular, inadequate software architectures were used: e.g., programs generally did not have any self-understanding. So although a programmer looking at program traces could detect searches stuck in loops, wasteful repetition, poor selection between options to explore first, and could notice opportunities for improving choice of representations, algorithms or repairing knowledge gaps, the programs could not do this to themselves. One of the few exceptions was the HACKER program reported in (Sussman, 1975). Even though it was not completely implemented, many of the issues were analysed by Sussman, and important aspects of meta-management are to be found both in his suggestion that 'critics' could monitor plan-formation processes by looking for instances of previously learnt bug-patterns, and also in HACKER's ability to record abstract features of processes of 'careful mode' plan execution, used for diagnosis of errors. But it is only recently that attention has been focused on architectures supporting a full range of interactions between concurrent processes.

It is widely believed that an emotional subsystem (whatever that might mean) is required to remove the deficits in earlier AI systems (Damasio, 1994, 1997). An alternative diagnosis is that human-like intelligence based on self-awareness and self-criticism requires a meta-management (reflective) layer (Minsky, 1987, 1994) in the architecture, operating concurrently with other components. Damage to frontal lobes in humans can interfere both with meta-management capabilities, leading to reduced intelligence, and with certain types of emotional reactions. The combined effects of one kind of damage have been misinterpreted by Damasio and others as implying that emotions are *required* for intelligence, as opposed to being a product of other things that are required for intelligence, as explained in (Sloman, 1998, 2001a). This is analogous to arguing that because damage to a car's battery will stop the horn working, and will also stop the car from starting, it follows that a working horn is required for the car to start.

The importance of a more sophisticated architectural approach is beginning to be widely acknowl-

edged – e.g., it plays a significant role in the recent DARPA initiative on cognitive systems, in the USA. [14] However, it would be a mistake to assume that the problem is solved, for we still have very little understanding of how to build powerful self-monitoring capabilities, or human multi-level perception and action systems, and we don't have good explanations for meta-semantic capabilities. We also still need to understand possible evolutionary and developmental trajectories in which architectures change.

Experiments illustrating the evolutionary impact of simple changes within the space defined by our framework are reported in (Scheutz & Sloman, 2001). There is much more exploration to be done of this sort and we have developed tools to help with the task (Sloman & Logan, 1999).

9.1. Using the CogAff framework to guide research

For decades AI has suffered from swings of fashion in which people have felt they have to propose and defend a particular type of architecture as 'right' and others as 'wrong'. This should be replaced by research that systematically compares design or modelling options, including hybrid combinations, in order to understand the trade-offs. The CogAff schema can be used to provide a framework that promotes research:

(1) asking questions about an organism: which of the sub-components and which of the links between components does it have, and what difference would it make if the architecture were different in various ways?
(2) asking similar questions about the ontologies and forms of representations used by different organisms, e.g., what are the affordances they can detect, and how can they use them?
(3) considering alternative designs for artificial systems, and investigating the pros and cons of including or omitting some of the sub-mechanisms or links between sub-mechanisms;
(4) illuminating evolutionary investigations by enabling us to identify and analyse possible evolutionary trajectories in design space and in niche space (Sloman, 2000a, 2001).
(5) challenging and extending the CogAff framework by noticing when useful proposed architectures do not fit the framework. For example, both Clarion and the H-CogAff architecture require kinds of complexity not directly suggested by the CogAff grid.

It is common in AI to argue that a particular sort of architecture is a good one and then to try to build instances in order to demonstrate its merits. However, this may be of limited value if it is not clear what the space of possible architectures is and what the trade-offs are between the different designs within that space. So even if a particular architecture supports some capability or produces some desired robot behaviour we may be left in the dark as to whether a different architecture might have explained more or produced a more useful or interesting robot.

Likewise performing evolutionary computation experiments to develop a good design to solve some problem will not increase our understanding of why one design works and another does not, unless we have a good ontology for describing designs and their relationships to various niches.

An *explicitly comparative* framework encourages a more analytical approach, even if it is not strictly necessary for finding good solutions to engineering problems. (Cf. discovering useful drugs for treating diseases, without understanding how the diseases or the drugs work.)

Comparative analysis of different designs also helps us understand how different architectures may produce the same behaviours, forcing us to develop more sophisticated criteria for evaluating scientific models of organisms than visible behavioural similarities.

10. Uncovering problems by designing solutions

Of course we are not proposing sitting in our armchairs and designing then implementing systems: any engineer knows that you only understand

[14] See http://www.darpa.mil/body/NewsItems/pdf/iptorelease.pdf and http://www.darpa.mil/ipto/.

the problem to be solved after you have designed and tested (including interacting with) a variety of prototype solutions. The same applies to understanding *what needs to be explained* in the case of natural information-processing systems. Often it is only when you discover surprising things a model does and does not do that you understand what its task specification should have been.

For example, a prototype may work as expected on a variety of planned test cases, then produce bizarre behaviour when a new test is attempted. Sometimes understanding this requires the researcher's ontology to be extended – for instance noticing that environments can differ in ways that were not previously noticed but are significant for the organism being modelled. An example could be discovering that the same perceptual information is interpreted differently in different contexts by the original organism but not by the prototype robot, pointing to the need for the robot to recognize and use information about such contexts. For instance, a robot car-driver will somehow have to acquire the ability to perceive intentions and plans of other drivers.

Another case is finding systems that work well when they can solve a problem but work very badly otherwise, for instance continuing to search blindly because they have not noticed that a strategy cannot succeed – like many AI systems and also the patient Eliot in (Damasio, 1994). That observation might draw attention to a previously unnoticed requirement for managing internal processing, as happened in research on symbolic AI problem-solving systems. Learning from partly unsuccessful prototypes has fuelled growth in concepts used by AI researchers over the last 50 years, including a switch from emphasis on representations and algorithms to emphasis on architectures (e.g., Sussman, 1975; Albus, 1981; Brooks, 1986; Minsky, 1987; Laird, Newell, & Rosenbloom, 1987; Beaudoin, 1994; Nilsson, 1998).

11. Asking questions about an information-processing system

Understanding an information-processing system requires us not only to find out how it behaves in various environments and how it is internally made up from physical components, but also to ask some questions about abstract features of its functionality. For example, if the architecture includes a deliberative layer we can ask what sorts of representational formalisms and mechanisms enable it to represent unperceived possibilities, and whether it uses one formalism for all contexts or different formalisms for different kinds of task. Can it describe relationships between hypothetical possibilities. Can it learn to invent new types of formalisms? To what extent does it require formalisms with varying structures and compositional semantics?

Similar questions can be asked about reactive systems or sub-systems, even though not everyone is happy to use the term 'representation' in that context. They may prefer to ask: What kinds of system states and processes can store information used by the system? How can those states vary (e.g., do they vary like vectors in a fixed dimensional vector space, or can they vary in structure and complexity like sentences or tree structures?). How is the information therein extended, compared, retrieved, and used, and for what functions? [15]

As said before, such questions refer to types of information and manipulative mechanisms that exist within *virtual* machines. So investigating them can be extremely difficult, since in general they cannot be observed using conventional scientific methods, either in externally observable behaviour nor in the physical or physiological processes in brains.

An example is the question: How do animals with deliberative capabilities represent collections of possibilities inherent in a situation? Modal logics can be used to represent possibilities and impossibilities, but it is not obvious that animal

[15] As in (Sloman, 1995, 1996b) we treat all these as questions about types of *representations*, their syntax, their semantics and their pragmatics, even if the representations are implemented in neural nets, chemical concentrations, patterns of wave activity, etc. Some theorists prefer to use 'representation' in a narrow way requiring a particular type of syntax (e.g., using phrase structure grammars) and semantics (e.g., with propositional content). This restriction seems pointless given the variety of types of information processing that would thereby be omitted.

minds use such formalisms (Sloman, 1996a). Likewise if a perceptual system detects affordances, we can ask whether this is implemented purely at the reactive level by triggering appropriate behavioural responses (as in insects and many other animals), or whether affordances are somehow described in a deliberative sub-system that can consider whether to make use of them and if so which ones, and how. A system with both layers might use both mechanisms in parallel, as Clarion does.

If there is a meta-management layer that includes meta-semantic self-monitoring and self-evaluation capabilities we can ask what sorts of categorisation of internal states are used, whether the evaluations are innate or learned, and, if learned, how much influence the surrounding culture has (e.g., whether individuals can feel guilt or shame, as opposed to merely regretting what they have done).

In summary: our general approach, and the CogAff schema in particular, leads to a wide range of new empirical and theoretical research questions.

12. Extending our design ontology

There may be some benefit to the community studying biologically-inspired robots if we generalise some of the currently used ontology, as has often happened in the history of science when new commonalities are discovered. For example, the label 'energy' was extended to entirely new phenomena such as chemical energy and mass energy, allowing a more general interpretation of the principle of conservation of energy, and the idea of 'feedback' was extended from mechanical controllers to electrical, chemical, biological and socio-economic processes.

Likewise instead of describing some systems as using representations and others as having changeable states that can store useful information, we can describe both as using representations, and then discuss the similarities and differences between the different types of representation and representation-manipulating mechanisms. We can then usefully extend the interpretation of questions like these:

- What kinds of syntax are used (what information structures and syntactic transformations)?
- What kinds of semantics are used (which ontologies, hypotheses, questions, explanations, beliefs, intentions, plans)?
- For what pragmatic functions is the information used (goals, desires, puzzles, strategies, preferences, values, triggering new states, etc.)?

Of course, not all these questions are relevant to all organisms. Since these phenomena are all very abstract, exploring them is not like perceiving physical behaviour, but requires us to develop appropriate meta-ontologies and new modes of investigation. But that is not unusual in science: similar developments were required before biologists could study abstractions like 'function', 'adaptation', 'metabolism', 'niche', 'gene' and 'extended phenotype'. Moreover the history of physics also includes major advances towards more abstract ontologies referring to entities that are remote from what is readily observed and measured. For example, Pauli postulated the existence of neutrinos in 1930, but 'detection' of neutrinos did not occur until about 25 years later, and even then required very elaborate inference procedures.

12.1. Enriching our conceptual frameworks

Our grasp of categories required for information processing in natural systems is still very limited, compared with the ontologies we have developed for designing and talking about artificial systems. In studying most animals we are probably in the situation of someone trying to understand what a computer is doing who has never studied operating systems, compilers, programming techniques, networking, word processors, data-bases, expert systems, etc.

A physicist or electronic engineer who knows nothing about these things may be able to investigate many of the physical and electronic properties of the computer, without ever dreaming that he is leaving anything out. Likewise it is possible for brain scientists to investigate in great detail physiological pathways, patterns of neural and chemical activity, and their correlations with external

events, and never dream that the investigation leaves out many important questions about the virtual machines involved, such as what ontology the organism uses in acquiring information about its environment, or what forms of syntax are used in storing and manipulating information of various kinds.

The ontology used by an organism will not be made visible by studying physical processes in its brain. Opening up the brain of an expert computer scientist will not teach you about compilers and schedulers. So in addition to the general possibility of ontological blindness: we may be ontologically blind to some aspect of the ontology used by an organism – second order ontological blindness.

Neither are the information-processing capabilities visible in the externally observable behaviour or input-output mappings displayed by machines or animals – except to those who have developed appropriate theories to guide their observations and interpretations. We therefore need ways of thinking about and investigating aspects of biological systems (organisms, species, ecosystems) that are not necessarily observable to us today, but may be crucial for understanding how they work. (We also need to teach these ways of thinking to more students.)

Understanding such (currently) 'invisible' aspects of processes in organisms may be a requirement for realistic simulations or models, especially when we are starting from inherently different physical implementations, such as computers and digital circuits instead of brains and neurones, or electro-mechanical devices instead of muscles and bones. People who do not address the questions regarding the important abstractions may therefore build simulations which very superficially model biological systems without realising that there are important phenomena to which they are ontologically blind and which they have not modelled.

In some cases, attempting to get desired results by trying to replicate physical structures using artificial components may fail, like early attempts to replicate bird flight; whereas replication at a higher level of abstraction may be more successful – as happened in the history of achieving artificial flight (Armer, 1962, p. 334) (reprinted in Chrisley, 2000).

13. Summary and conclusion

The ability of organisms (whether seagulls or scientists) to perceive and reason about the world obviously means that we can and should allow for this possibility when thinking about how to design biologically-inspired robots. What is not so clear is what precisely the perception and reasoning capabilities of particular types of animal are, and that includes what forms of representation they use. Up to a point this can be investigated by performing standard engineering requirements analysis in order to work out how to replicate or model observed capabilities. Unfortunately for researchers, information-processing systems not only produce easily observable physical actions, but also have states such as perception, reasoning, learning, desires, and emotions, with aspects that are difficult or impossible to observe using the methods of the physical sciences or even our evolutionarily-honed abilities to see mental states of others. Thus an observation-based approach to biologically inspired robots may miss important phenomena.

The abstraction of a virtual information-processing machine, itself understandable on various layers of abstraction, seems to be required to explain many biological behaviours. In particular some biologically-inspired robots should include not just reactive but also deliberative and meta-management layers, not for only the obvious reason that mathematicians do these things and they are biological entities, but also because merely reactive architectures seem to be unable to explain capabilities of many vertebrates, for instance the behaviour of a crow in bending a piece of wire to make a hook to fish a bucket of food out of a tall tube. [16] Animals that not only introspect but also report their introspections (not necessarily accurate or complete introspections) are likely to

[16] Reported in August 2002, in several newspapers, news web sites and journals, e.g., here: http://news.nationalgeographic.com/news/2002/08/0808_020808_crow.html The process could not be purely reactive unless something in the crow's evolutionary or individual history produced either genetic or learnt hook-making reactions. There does not appear to have been any such evolutionary history or prior training of the individual crow.

be using something like the CogAff meta-management layer (unless, like a parrot repeating what it hears, their reports are faked). Since not many animals can report introspections, evidence for meta-management will have to be very indirect.

A desire for theoretical parsimony, or unfounded worries concerning scientific respectability of speculation about processes which are unobservable or difficult to observe, may help to preserve ontological blindness, by causing some researchers to adopt a methodology that permits only low-level physical phenomena (i.e., phenomena describable using the language of the physical sciences) to play a role in explanations and designs. This 'narrow' viewpoint rules out using our draft schema for possible architectures and broadening it to encompass all the forms that biological cognition might take, so that we can investigate architectures with a more open mind. This should be a useful counter to some recent restrictive influences, such as the ideas presented in (Brooks, 1986) leading to so-called 'Nouveau AI'.

We have tried to show how the CogAff framework accommodates many architectures proposed so far, including subsumption, varieties of contention-scheduling, and other 'Omega' architectures, Barkley's 'executive functions', aspects of Clarion, and others. Even if the precise schema we have proposed proves insufficiently general, there will still be a need for something like it as a unifying framework for AI, theoretical psychology and neuroscience. A demanding test for the ideas in this paper could come out of attempts to build a child-like robot with a great deal of the visual capability, the physical manipulative capability, the linguistic capability, the ability to use and to provide explanations, and the capability to learn and develop, of an 'idealised' young human child.

Acknowledgements

This work was funded by the Leverhulme Trust, and is based in part on earlier collaboration with Brian Logan and Matthias Scheutz, funded by the Trust. Luc Beaudoin's PhD thesis made a major contribution to our thinking about architectures. Further details can be found in papers at this site: http://www.cs.bham.ac.uk/research/cogaff/. Our software tools are available at this site: http://www.cs.bham.ac.uk/research/poplog/. Minsky's draft book *The Emotion machine*, is very relevant. It can be found at his home page: http://www.media.mit.edu/people/minsky/. A toolkit designed to support exploration of architectures within the CogAff framework can be found here: http://www.cs.bham.ac.uk/research/poplog/packages/simagent.html. A new EC-funded 'cognitive systems' project will extend many of the ideas developed here, as described in http://www.cs.bham.ac.uk/research/projects/cosy/. We are grateful for useful comments from referees and the editor.

References

Albus, J. (1981). *Brains, behaviour and robotics. Byte books*. Peterborough, NH: McGraw-Hill.

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

Armer, P. (1962). Attitudes toward intelligent machines. In E. Feigenbaum & J. Feldman (Eds.), *Computers and Thought*. New York: McGraw-Hill, Reprinted in (Chrisley, 2000).

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

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* (p. 238). Amsterdam: IOS Press.

Boden, M. A. (2000). Autopoesis and life. *Cognitive Science Quarterly, 1*(1), 115–143.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: The MIT Press.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation, RA-2*, 14–23, 1.

Carnap, R. (1947). *Meaning and necessity: a study in semantics and modal logic*. Chicago: Chicago University Press.

Chomksy, N. (1959). Review of skinner's Verbal Behaviour. *Language, 35*, 26–58.

Chrisley, R. (2000). *Artificial intelligence: Critical concepts volume 1*. London: Routledge.

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

Damasio, A. (1994). *Descartes' error, emotion reason and the human brain*. New York: Grosset/Putnam Books.

Dawkins, R. (1982). *The Extended Phenotype: The long reach of the gene*. Oxford, New York: Oxford University Press.

Evans, T. (1968). A heuristic program to solve geometric analogy programs. In M. Minsky (Ed.), *Semantic Information Processing* (pp. 271–353). Cambridge, MA: MIT Press.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gibson, J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates (originally published in 1979).

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

Hauser, M. (2001). *Wild minds: What animals really think*. London: Penguin.

Karmiloff-Smith, A. (1996). Internal representations and external notations: a developmental perspective, in (Peterson, 1996), pp. 141–151.

Kim, J. (1998). *Mind in a physical world*. Cambridge, MA: MIT Press.

Kohler, W. (1927). *The mentality of apes* (2nd ed.). London: Routledge & Kegan Paul.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence, 33*, 1–64.

Lakatos, I. (1970). *Criticism and the growth of knowledge*. New York: Cambridge University Press.

Lovelock, J. (1979). *Gaia: A new look at life on earth*. Oxford: Oxford University Press.

Marr, D. (1982). Vision. Freeman.

McCarthy, J. (1979). Ascribing mental qualities to machines. In M. Ringle (Ed.), *Philosophical perspectives in artificial intelligence* (pp. 161–195). Atlantic Highlands, NJ: Humanities Press, Also accessible athttp://www-formal.stanford.edu/jmc/ascribing/ascribing.html .

Minsky, M., Singh, P., & Sloman, A. (2004). The St. Thomas common sense symposium: designing architectures for human-level intelligence. AI Magazine, 25(2), 113-114. Available: http://web.media.mit.edu/~push/StThomas-AIMag.pdf.

Minsky, M. L. (1987). *The society of mind*. London: William Heinemann Ltd.

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*. San Francisco: Morgan Kaufmann.

Norman, D., Ortony, A., & Russell, D. (2003). Affect and machine design: Lessons for the development of autonomous machines. *IBM Systems Journal, 42*, 38–44, http://www.research.ibm.com/journal/sj/421/norman.pdf.

Picard, R. (1997). *Affective computing*. Cambridge, MA, London, England: MIT Press.

Rose, S. (1993). *The making of memory*. Toronto, London, New York: Bantam Books.

Scheutz, M. (2001). The evolution of simple affective states in multi-agent environments. In D. Cañamero (Ed.), *Proceedings AAAI fall symposium 01* (pp. 123–128). Falmouth, MA: AAAI Press.

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

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

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

Sloman, A. (1982). Towards a grammar of emotions. *New Universities Quarterly, 36* (3) 230–238. http://www.cs.bham.ac.uk/research/cogaff/0-INDEX96-99.html#47.

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, Available at http://www.cs.bham.ac.uk/research/cogaff/.

Sloman, A. (1993). The mind as a control system. In C. Hookway & D. Peterson (Eds.), *Philosophy and the cognitive sciences* (pp. 69–110). Cambridge, UK: Cambridge University Press.

Sloman, A. (1995). Musings on the roles of logical and non-logical representations in intelligence. In J. Glasgow, H. Narayanan, & B. Chandrasekaran (Eds.), *Diagrammatic reasoning: Computational and cognitive perspectives* (pp. 7–33). Cambridge, MA: MIT Press.

Sloman, A. (1996a). Actual possibilities. In L. Aiello & S. Shapiro (Eds.), *Principles of knowledge representation and reasoning: proceedings of the fifth international conference (KR '96)* (pp. 627–638). Boston, MA: Morgan Kaufmann Publishers.

Sloman, A. (1996b). Towards a general theory of representations. In D. M. Peterson (Ed.), *Forms of representation: an interdisciplinary theme for cognitive science* (pp. 118–140). Exeter, UK: Intellect Books.

Sloman, A. (1998). Damasio, Descartes, alarms and meta-management. In *Proceedings international conference on systems, man, and cybernetics (SMC98), San Diego* (pp. 2652–2857). New York: IEEE.

Sloman, A. (2000a). Interacting trajectories in design space and niche space: A philosopher speculates about evolution. In M. Schoenauer et al. (Eds.), *Parallel problem solving from nature – PPSN VI. Lecture Notes in Computer Science* (No 1917, pp. 3–16). Springer-Verlag: Berlin.

Sloman, A. (2000b). Models of models of mind. In M. Lee (Ed.), *Proceedings of symposium on how to design a functioning mind, AISB'00* (pp. 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 T. Cootes & C. Taylor (Eds.), *Proceedings*

of british machine vision conference (pp. 313–322). Manchester: BMVA.

Sloman, A. (2002a). How many separately evolved emotional beasties live within us?. In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in Humans and Artifacts* (pp. 35–114). Cambridge, MA: MIT Press.

Sloman, A. (2002b). The irrelevance of Turing machines to AI. In M. Scheutz (Ed.), *Computationalism: New Directions* (pp. 87–127). Cambridge, MA: MIT Press, Available at http://www.cs.bham.ac.uk/research/cogaff/.

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

Sloman, A., & Logan, B. (1999). Building cognitively rich agents using the Sim_agent toolkit. *Communications of the Association for Computing Machinery, 42*(3), 71–77.

Sloman, A., & Scheutz, M. (2001). Tutorial on philosophical foundations: Some key questions. In *Proceedings IJCAI-01* (pp. 1–133), Menlo Park, CA: AAAI. http://www.cs.bham.ac.uk/~axs/ijcai01.

Sloman, A., Chrisley, R., & Scheutz, M. (2004). The architectural basis of affective states and processes. In M. Arbib & J.-M. Fellous (Eds.), *Who Needs Emotions?: The Brain Meets the Machine*. Oxford, New York: Oxford University Press, Online athttp://www.cs.bham.ac.uk/research/cogaff/sloman-chrisley-scheutz-emotions.pdf .

Sun, R. (2002). *Duality of the mind*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

von Neumann, J. (1951). The general and logical theory of automata. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium*. Hafner Publishing. Reprinted in (Chrisley, 2000).

Weiskrantz, L. (1997). *Consciousness lost and found*. New York, Oxford: Oxford University Press.

Wiener, N. (1961). *Cybernetics: or control and communication in the animal and the machine* (2nd ed.). Cambridge, MA: The MIT Press.

Zaera, N., Cliff, D., & Bruten, J. (1996). (Not) Evolving Collective Behaviors in Synthetic Fish. In P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, J., & Wilson, S. (Eds.). *From animals to animats 4: Proceedings of the fourth international conference on simulation of adaptive behavior (SAB96)* (pp. 635–644). MIT Press, Bradford Books.