Architectures and Tools for Human-Like Agents

Aaron Sloman and **Brian Logan**
School of Computer Science
The University of Birmingham
Birmingham, B15 2TT, UK
+44 121 414 {4775 (Sloman) 3712 (Logan)}
{A.Sloman, B.S.Logan}@cs.bham.ac.uk

ABSTRACT[1]
This paper discusses agent architectures which are describable in terms of the "higher level" mental concepts applicable to human beings, e.g. "believes", "desires", "intends" and "feels". We conjecture that such concepts are grounded in a type of information processing architecture, and not simply in observable behaviour nor in Newell's knowledge-level concepts, nor Dennett's "intentional stance." A strategy for conceptual exploration of architectures in design-space and niche-space is outlined, including an analysis of design trade-offs. The SIM_AGENT toolkit, developed to support such exploration, including hybrid architectures, is described briefly.
Keywords:
Architecture, hybrid, mind, emotion, evolution, toolkit.

MENTALISTIC DESCRIPTIONS

The usual motivation for studying architectures is to explain or replicate performance. Another, less common reason, is to account for concepts. This paper discusses "high level" architectures which can provide a systematic non-behavioural conceptual framework for mentality (including emotional states). This provides a new kind of semantics for mentalistic descriptions. We illustrate this using multi-layered architectures based in part on evolutionary considerations. We show briefly how different layers support different sorts of emotion concepts. This complements work by McCarthy(1979, 1995) on descriptive and notational requirements for intelligent robots with self-consciousness.

We provide pointers to an uncommitted software toolkit that supports exploration of hybrid architectures of various sorts, and we illustrate some of the architectural complexity it needs to support.

WHY USE MENTALISTIC LANGUAGE?

We shall need mentalistic descriptions for artificial agents for the same reasons as we need them for biological agents, e.g. (a) because such descriptions will (in some cases) be found irresistible and (b) because no other vocabulary will be as useful for describing, explaining, predicting capabilities and behaviour. ((b) provides part of the explanation for (a).) So, instead of the self-defeating strategy of trying to avoid mentalistic language, we need a disciplined approach to its use, basic mentalistic concepts on information-level architectural concepts.

The "information level" design stance

Dennett (1978) recommends the "intentional stance" in describing sophisticated robots, as well as human beings. That restricts mentalistic language to descriptions of whole agents, and presupposes that the agents are largely rational. Similarly, Newell (1982) recommends the use of the "knowledge level", which also presupposes rationality. By contrast, we claim that mentality is primarily concerned with an "information level" architecture, close to the requirements specified by software engineers. This extends Dennett's "design stance" by using a level of description between physical levels (including physical design levels) and "holistic" intentional descriptions.

"Information level" design descriptions allow us to refer to various *internal* semantically rich short term and long term information structures and processes. This includes short term sensory buffers, longer term stored associations, generalisations about the environment and the agent, stored information about the local environment, currently active motives, motive generators that can produce motives under various conditions, mechanisms and rules for detecting and resolving conflicts, learnt automatic responses, mechanisms for constructing new plans, previously constructed plans or plan schemata, high level control states which can modulate the behaviour of other mechanisms, and many more.

Some mentalistic concepts refer to the information processing and control functions of the architecture. These functions include having and using information *about* things. E.g. an operating system has and uses information *about* the processes it is running. Here semantic content is present without full-blown intentionality or rationality. Restricting semantic notions to global states of a rational

Copyright © 2010, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.
[1]In European Conference on Cognitive Modelling, Nottingham, April 1998

agent, or banning them altogether from explanatory theories, would be as crippling in the study of intelligent agents as it would be in the engineering design of complex control systems. (However, not all semantic states can be fully characterised in terms of *internal* functions, for instance those that refer to *particular* external objects, such as Buckingham Palace, a point beyond the scope of this paper.)

Many of the mechanisms in such an architecture are neither rational nor irrational: even though they acquire information, evaluate it, use it, store it, etc. (Sloman 1994*b*). They are neither rational nor irrational because they are *automatic*. Even a deliberative architecture at some level needs reactive mechanisms to drive the processing. If everything had to be based on prior goals and justifications nothing would ever happen.

ARCHITECTURAL ANALYSIS

Different architectures can correspond to different views of a system, e.g. a physical architecture, composed of the major physical parts, a physiological architecture, corresponding to the major functional roles of physical parts, and an information processing architecture composed of mechanisms involved in acquiring, transforming, storing, transmitting, and using information.

There need not be a one to one correspondence between components in different views. A physical component may be shared between several physiological functions: e.g. the circulatory system is involved in distribution of energy, waste disposal, temperature control, and information transfer.

There is a huge space of possible designs. We make no presumption that information processing mechanisms must all be computational (whatever that means). Nor is there a commitment regarding *forms* used to encode or express information. They may include logical databases, procedures encoding practical know-how, image structures, neural nets or even direct physical representations, as in thermostats and speed governors.

Biological plausibility requires evolvability as well as consistency with experimental data and brain physiology. The capabilities and neural structures of different sorts of animals (e.g. insects, rodents, apes, humans) suggest that different types of architectures evolved at different times, with newer architectures building new sorts of functionality on older ones. We suggest that human mental states and processes depend on interactions between old and new layers in a biologically plausible control architecture producing various kinds of internal and external behaviour, including "internal" processes such as motive generation, attention switching, global redirection in emergencies, problem solving, information storage, skill acquisition, self-evaluation and even modification of the architecture.

Besides the multi-layered central information processing architecture there are sensors and effectors of various kinds. These involve more than just transduction of energy or information into or out of the system. We suggest that both have evolved multiple layers interacting with the different layers in the central system as in Figure 1. Such an architecture can generate a huge variety of concepts relevant to describing its states and processes. It also supports a wide variety of types of learning, yet to be analysed.

Indeterminacy of architecture

Often boundaries between sub-mechanisms and levels of description are unclear, including the boundary between the control architecture and mere physiological infrastructure. In brains, chemical processes provide energy and other resources, along with damage repair and resistance to infections. However, effects of drugs, diseases and genetic defects involving brain chemicals suggest that chemistry forms more than a physiological infrastructure: chemically controlled mood changes may be an important part of an organism's intelligent reaction to changing circumstances, and alcohol can change "no" into "yes"! But we don't know how far chemical reactions play a direct role in information processing or high level control,

In both perception and action the "hardware/software" boundary is blurred. E.g. visual attention can be switched with or without redirection of gaze, and fine-grained manipulation can be shared between software and hardware, e.g. in compliant wrists, which reduce the control problem in pushing a close fitting cylinder into a hole. Simon (1969) pointed out long ago that there can be information sharing between internal and external structures.

It is too early for clear definitions of the boundaries of architectures or their components. However, important ideas are beginning to emerge including contrasts between:
(a) reactive *vs* deliberative functions,
(b) symbolic *vs* neural mechanisms,
(c) logical *vs* other sorts of information manipulation,
(d) continuous *vs* discrete control,
(e) using continuously available environmental information *vs* using information stored in memory,
(f) hierarchical *vs* distributed control,
(g) serial *vs* concurrent processing,
(h) synchronised *vs* asynchronous processing,
(i) genetically determined capabilities, those produced by adaptive mechanisms within individuals, and those absorbed from a culture (e.g. learnt poems and equations).

Instead of viewing these contrasts as specifying *rival* options, we should allow combinations of these alternatives to have roles in multifunctional architectures. Work on hybrid mechanisms (e.g. combinations of neural and symbolic systems) is now commonplace, but in order to explore agents rivalling human or even chimpanzee sophistication we need to understand far more complex combinations of subsystems, including complex sub-architectures *within* perceptual and motor control mechanisms, and a deep integration of cognitive

and affective functions and mechanisms (Wright, Sloman & Beaudoin 1996, Sloman 1998(forthcoming)). However, there is no unique "correct" architecture: different designs have different trade-offs, as biological evolution shows. We need to understand the trade-offs and possible trajectories. This includes finding good concepts for describing systems with different designs.

ARCHITECTURES AND EMERGENT CONCEPTS

A deep conceptual framework takes account of the range of possible states and processes supported in an architecture, generating a system of high-level descriptive concepts for describing an organism, software agent, or robot, just as a knowledge of molecular architecture provides a basis for labelling chemical compounds and describing chemical processes.

A control architecture can support a collection of states and processes, often indefinitely large. Concepts derived in this way from the architecture are "deep concepts". "Shallow" concepts, based entirely on observed behavioural patterns bearing no relationship to the architecture, are likely to have reduced predictive and explanatory power, like concepts of physical matter based on visible properties rather than atomic and molecular structure.

Not all states require specific mechanisms in the architecture. A computing system that is "overloaded" does not have an "overloading" mechanism, since overloading results from interaction of many different mechanisms whose functions is not to produce overload. Similarly many mental states, e.g. some debilitating emotions, may *emerge* from interactions within an architecture, rather than from an emotion module.

If there are several coexisting, interacting sub-architectures (e.g. reactive and deliberative sub-architectures) then higher order concepts are needed to describe the variety of possible relationships between them. For instance, states in one subsystem can modulate processes in others. Such relationships can change over time: sometimes one part is dominant and sometimes the other. Moreover, when training increases fluency in a cognitive skill this may shift responsibility for a task from a general purpose module to a dedicated module.

Familiar prescientific concepts, e.g. "emotion", can be ambiguous if they sometimes refer to processes in a component of the architecture (e.g. being startled, or terrified by a fast approaching menace, may result from a specific module, perhaps part of the limbic system) and sometimes to emergent interactions between subsystems (e.g. guilt and self-reproach).

Unlike emotions which we share with rats, e.g. being startled, which use this old global alarm system, many human emotions involve a partial loss of control of thought processes, (e.g. extreme grief, ecstasy or hysteria). This presupposes the possibility of being in control. That, in turn, depends on the existence of an architecture that supports certain kinds of self monitoring, self evaluation, and self modulation. Being careful or careless requires an architecture able to control which checks are made during planning, deciding and acting.

Which animal architectures can support control of thought processes is not clear. Systems lacking such underpinnings may not be usefully describable as "restrained", "resisting temptation", etc. Can a rat sometimes control and sometimes lose control of its thought processes? Can a rat be careless in its deliberations? Over-simple architectures in software agents will also make such concepts inappropriate to them.

EVOLUTION AND MODULARITY

Our discussion has presupposed that architectures are to some extent intelligible. Will naturally evolved systems be modular and intelligible? In principle, any required finite behaviour could be produced by a genetically determined, unstructured, non-modular architecture, including myriad shallow condition-action rules with very specific conditions and actions providing flexibility. However, as the diversity of contexts grows and the need to cope with unexpected situations, including interactions with other other agents, increases, memory requirements for such a system can grow explosively, and it becomes more difficult find a design which anticipates all the conditions and actions in advance. Thus the time required to evolve all the shallow capabilities is far greater and the required diversity of evolutionary contexts far greater than for a system with planning abilities.

A shallow non-modular system would not only be hard to design, describe and explain: it would be hard to control or modify, whether controlled from outside or controlling itself, whether modified by a designer, or modified by evolution. (Contrast the use of bit-strings in genetic algorithms with the use of trees in genetic programming.)

All this suggests that for complex organisms there would be pressure towards more modular architectures with generic mechanisms that can be combined by a planner to handle new situations, and adaptive architectures that can change themselves to improve performance. Both the normal evolutionary pressures for modularity and reuse, and the need for economy in high level self-control mechanisms could have increased the pressure towards evolution of modular control architectures, in some organisms. So the existence of self-monitoring, self-evaluation and self-control processes could influence the further evolution of control architectures. Apparently insects found a different solution.

It may eventually be possible to investigate this issue in simulated evolution.

THE EMERGENCE OF "QUALIA"

If a system has the ability to monitor its own states and processes, a new variety of descriptions becomes applicable, labelling new forms of self control, including its own discovery of concepts for self-description. The objects of such self-monitoring processes may be virtual machine states as well as internal physical or physiological

states.

Many of the spatial, temporal and causal categories used in perceiving the environment have evolved to support biological functions of organisms in those environments, even though precise details can vary widely between species and between individuals in a species. Likewise, it is possible that the basic and most general mentalistic categories that humans use in describing and thinking about themselves and other agents are not reinvented by different individuals (or cultures) but generated by evolutionary processes driving development of self-monitoring capabilities.

Phenomena described by philosophers as "qualia" may be explained in terms of high level control mechanisms with the ability to switch attention from things in the environment to *internal* states and processes, including intermediate sensory datastructures in layered perceptual systems. These introspective mechanisms may explain a child's ability to describe the location and quality of its pain to its mother, or an artist's ability to depict how things look (as opposed to how they are). Software agents able to inform us (or other artificial agents) about their own internal states and processes may need similar architectural underpinnings for qualia.

From this standpoint, the evolution of qualia would not be a single event, but would involve a number of steps as more kinds of internal states and processes became accessible to more and more kinds of self-monitoring processes with different functions, e.g. requesting help from others or discovering useful generalisations about oneself. Such step-wise development may also occur within an individual.

HOW TO MAKE PROGRESS

There are several ways in which we might try to explore the relationship between architecture and mentality. One approach is to push the approach based on "shallow" behaviour-based concepts as far as possible, and analyse where it breaks down, or where patching it is very difficult (e.g. dealing with new unexpected combinations of conditions where applicable rules conflict, or where no rule applies).

Another approach is to attempt a theoretical analysis of the types of situations that will make development increasingly difficult and to produce increasingly general architectures to cope with the difficulties, using any ideas that work, and then conducting experiments to find out where they break down. This approach need not be constrained by theories of how human minds work: there may be alternative architectures capable of producing extremely useful or even "believable" performances. Initially the constraints on this type of theorising will be very ill-defined because of paucity of relevant knowledge and the shallowness of current theories. However, it is likely that as the work progresses more and more constraints can come from advances in other fields, and more and more tests can be generated to help us choose between alternative hypotheses. (Compare the ancient Greek atomic theory with modern atomic theory.)

Yet another approach is to use whatever direct or indirect evidence is available from brain science, experimental psychology, forms of mental disorder, patterns of development in infancy and decay in old age, evolution, folklore, introspection, common observation, or conceptual analysis of everyday mental concepts. Plausible architectures based on such evidence can then be tested by running experimental implementations, or by analysing their consequences and performing empirical research.

Our work is based on the second and third approaches. The architectural ideas in this paper come from a wide range of sources.

ARCHITECTURAL LAYERS

Part of the task is to find increasingly accurate and explicit theories of the types of architecture to be found in various sorts of human minds (and others) to be used as frameworks for generating families of descriptive concepts applicable to different sorts of humans (including infants and people with various kinds of brain damage) and different sorts of animals and artificial agents.

We conjecture that human-like agents with powers of self-control need a type of architecture with at least three distinct classes of mechanisms which evolved at different times (Sloman 1998(forthcoming)):
(1) Very old reactive mechanisms, found in various forms in all animals, including insects — this includes "routine" reactive mechanisms and "global alarm" mechanisms (the limbic system).
(2) More recently evolved deliberative mechanisms, found in varying degrees of sophistication in some other animals (e.g. cats, monkeys);
(3) An even more recent meta-management (reflective) layer providing self-monitoring self-evaluation, and self-control, using in part deliberative mechanisms of type (2), and perhaps found only in humans and other primates (in simpler forms).

Such an architecture is shown schematically (without alarms) in Figure 1 and each of the layers is described in more detail below. Note that the layers occur in perceptual and motor subsystems as well as centrally.

This is one among many possible designs. Some animals or artefacts may have only one or two layers, and different kinds of reactive, deliberative and meta-management mechanisms are possible.

We are not claiming that these mechanisms are alike in all humans. Deliberative capabilities seem very primitive in new born infants, and the third layer may be non-existent at birth. Moreover a culture can influence development of these layers, as can effects of brain damage, disease or aging. Some architectures may be possible for synthetic agents that are never found in organisms (e.g. solely deliberative architectures, or hybrid systems without

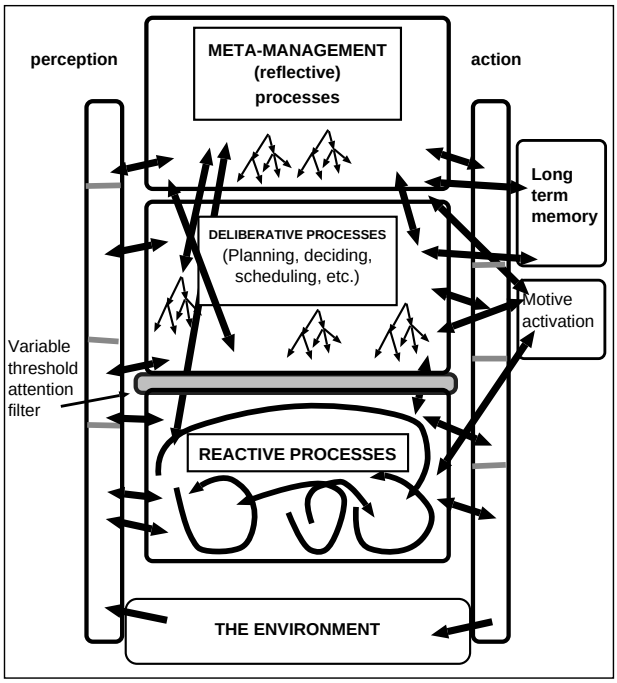

Figure 1: **A three layered agent Architecture**
(Note: global 'alarm' mechanisms not shown.)

global alarms).

Categories and strategies in all layers may be influenced by physical and social environments. A meta-management layer may use both categories and values absorbed from a culture as well as some genetically determined categories and strategies. For instance, certain motives for acting promote negative self-assessment and guilt in some cultures and not in others.

Within an individual, it is also possible for different modes of meta-management to take control in different contexts, e.g. in a family context, in a football game, and in the office. Individual variations might lead, at one extreme to multiple-personality disorder, and at another extreme to excessively rigid personalities.

Concurrent mechanisms

The layers are not assumed to form a rigidly hierarchical control architecture. Rather the three layers operate concurrently, with mutual influences. The reactive mechanisms will perform routine tasks using genetically determined or previously learnt strategies. When they cannot cope, deliberative mechanisms may be invoked, by the explicit generation of goals to be achieved. This can trigger various kinds of deliberative processes including considering whether to adopt the goal, evaluating its importance or urgency, working out how to achieve it, comparing it with other goals, deciding when to achieve it, deciding whether this requires reconsideration of other goals and plans, etc. (See chapter 6 of Sloman (1978).)

At other times the deliberative mechanisms may either attend to long term unfinished business or run in a "free-wheeling" mode, nudged by reactive processes which normally have low priority, including attention-diverting mechanisms in the perceptual subsystems. To allow direct communication with "higher" cognitive functions, perceptual systems may also have layered architectures in which different levels of processing occur in parallel, with a mixture of top-down and bottom-up processing. (Compare seeing a face as a face and as happy.)

If the internal layers operate concurrently, fed in part by sensory mechanisms which are also layered, they may also benefit from a layered architecture in motor systems. For example, reactive mechanisms may directly control some external behaviour, such as running, while the other mechanisms are capable of modulating that behaviour (e.g. changing the speed or style of running, or in extreme cases turning running into dancing). Likewise proprioceptive feedback of different sorts may go to different layers.

Where there is a global alarm system, there may be variations as regards which components provide its inputs and which can be modified by it. In humans connections to and from the limbic system seem to exist everywhere (Goleman 1996).

We now describe in a little more detail the differences between the layers (Figure 1) before discussing their implications for emotions. (The figure is much simplified, to reduce clutter).

Reactive agents

It is possible for an agent to have a purely reactive architecture, where:

- Mechanisms and space are permanently dedicated to specific functions, and can run concurrently, more or less independently, with consequent speed benefits. Some may be digital, some continuous.
- Conflicts may be handled by vector addition, voting, or winner-takes-all nets.
- Some learning is possible: e.g. tunable control loops, change of weights by reinforcement learning. Such learning merely alters links between pre-existing structures and behaviours.
- There is no explicit construction of new plans or structural descriptions or other complex internal objects, and therefore no explicit evaluation of alternative structures.
- Concurrent processing at different abstraction levels can encourage the evolution of different levels of processing in sensory and motor subsystems.
- Some of the reactions to external or internal conditions may be internal, e.g. various kinds of internal feedback control loops.
- If "routine" reactions are too slow a fast "global alarm" system taking control in emergencies may be useful.

As explained above, if all the main possible behaviours need to be built in by evolutionary adaptation or direct

programming the space requirements may explode as combinations increase. Likewise the time required to evolve all relevant combinations. A partial solution is to provide "chaining" mechanisms so that simpler behaviours can be re-used in different longer sequences. Simple sub-goaling may achieve this, changing internal conditions that launch behaviours. This may be a precursor to deliberative mechanisms.

It appears that insects have purely reactive architectures, and cannot reflect on possible future actions. Yet the reactive behaviours can produce and maintain amazing construction, e.g. termites' "cathedrals".

There is no form of externally observable behaviour that cannot, in principle, be implemented in a purely reactive system, without any deliberative capabilities, though it seems that in some organisms the evolutionary pressures mentioned above have led towards a different solution — which may coexist with the old one.

Combining reactive and deliberative layers

The ability to construct new complex behaviours as required reduces the amount of genetic information that needs to be transmitted as well as the storage requirements for each individual. It also reduces the number of generations of evolution required to reach a certain range of competence. In a deliberative mechanism:

- Evaluating and comparing options for novel combinations before selecting them requires a new ability to build internal descriptions of internal structures. It also needs a long term associative memory.

- Using re-usable storage space for new plans and other temporary structures, and use of a single associative memory (even if based on neural nets), makes processes inherently serial.

- New behaviours developed by the deliberative system can be transferred to the reactive layer (e.g. learning new fluent skills).

- Sensory and action mechanisms may develop new, more abstract, processing layers, which communicate directly with deliberative mechanisms. This could explain high level sensory experiences (e.g. seeing a face as happy).

- Even if neural nets are used, operation may be resource-limited because learning from consequences becomes explosive if too many things are done in parallel. Limiting concurrent processes may also simplify integrated control.

- Deliberative resource limits may mean that a fast-changing environment can cause too many interrupts and re-directions. Filtering new interrupts via dynamically varying thresholds (see Figure 1) helps but does not solve all problems.

- A global alarm system may include inputs from and outputs to deliberative layers.

The need for self-monitoring (meta-management)

Deliberative mechanisms may be implemented in specialised reactive mechanisms which react to internal structures, and can interpret explicit rules and plans.

However, evolutionarily determined deliberative strategies for planning, problem solving, decision making, evaluating options, can be too rigid. Internal monitoring mechanisms may help to overcome this e.g. by recording deliberative processes and noticing which planning strategies or attention switching strategies work well in which conditions. This could include detecting when one goal is about to interfere with other goals, or noticing that a problem solving process is "stuck", e.g. in a loop, or noticing that a solution to one problem helps with another.

Internal monitoring combined with learning mechanisms may allow discovery of new ways of categorising internal states and processes and better ways of organising deliberation. Meta-management and deliberative mechanisms permit cultural influences via the absorption of new concepts and rules for self-categorisation, evaluation and control.

Attending to intermediate perceptual structures can also allow more effective communication about external objects, e.g. by using viewpoint-centred appearances to help direct attention, or using drawings and paintings to communicate about how things look.

The meta-management layer may share mechanisms with the other two, including the global alarm mechanism (limbic system?) but also needs new mechanisms that can access states and processes in various parts of the whole system, categorise what is going on internally, evaluate it, and in some cases modify it. This can help with proper management of limited deliberative resources.

ARCHITECTURAL LAYERS & EMOTION CONCEPTS

We conjecture that different layers account for different sorts of mental states and processes, including emotional states. Disagreements about the nature of emotions can arise from failure to see how different concepts of emotionality depend on different architectural features, not all shared by all the animals studied.

(1) The old reactive layer, with the global alarm system, produces rapid automatically stimulated emotional states found in many animals (being startled, terrified, sexually excited).

(2) A deliberative layer, in which plans can be created and executed, supports cognitively rich emotional states linked to current desires plans and beliefs (like being anxious, apprehensive, relieved, pleasantly surprised).

(3) Characteristically human emotional states (e.g. humiliation, guilt, infatuation, excited anticipation) can involve reduced ability to focus attention on important tasks because of reactive processes (including alarm processes) interrupting and diverting deliberative mechanisms, sometimes conflicting with meta-management decisions (Wright et al. 1996).

The second class of states depends on abilities possessed by fewer animals than those that have reactive capabilities. The architectural underpinnings for the third class are

relatively rare: perhaps only a few primates have them.

Many theories of emotion postulate a system that operates in parallel with normal function and can react to abnormal occurrences by generating some kind of interrupt, like the global alarm mechanism. Consider an insect-like organism with a purely reactive architecture, which processes sensory input and engages in a variety of routine tasks (hunting, feeding, nest building, mating, etc.). It may be useful to detect certain patterns which imply an *urgent* need to react to danger or opportunity by freezing, or fleeing, or attacking, or protecting young, or increasing general alertness. Aspects of the limbic system in vertebrate brains seem to have this sort of function (Goleman 1996).

In architectures combining reactive and deliberative layers, the alarm mechanism can be extended to cause sudden changes also in *internal* behaviour, such as aborting planning or plan execution, switching attention to a new task, generating high priority goals (e.g. to escape, or to check source of a noise). Likewise processing patterns in the deliberative layer may be detected and fed into the alarm system, so that noticing a risk in a planned action can trigger an alarm.

Where a meta-management layer exists, data from it could also feed into the alarm system, and it too could be affected by global alarm signals. One meta-management function could involve learning which alarm signals to ignore or suppress. Another would extend the alarm system to react to new patterns, both internal and external. Another would be development of more effective and more focused (less global) high speed reactions, e.g. replacing a general startle reaction with the reactions of a highly trained tennis player.

This, admittedly still sketchy, architecture, explains how much argumentation about emotions is at cross-purposes, because people unwittingly refer to different sorts of mechanisms which are not mutually exclusive. An architecture-based set of concepts can be made far less ambiguous.

Familiar categories for describing mental states and processes (e.g. believes, desires, perceives, attends, decides, feels, etc.) may not survive unchanged as our knowledge of the underlying architecture deepens, just as our categories of kinds of physical stuff were refined after the development of a new theory of the architecture of matter. Researchers need to be sensitive to the relationships between pre-theoretical and architecture-based concepts as illustrated in (Wright et al. 1996).

THE SIM_AGENT TOOLKIT

We still have much to learn about different agent architectures. The properties of complex systems cannot all be determined by logical and mathematical analysis: there is a need for a great deal more exploration of various types of architectures, both in physical robots and in simulated systems.

Many robot laboratories are doing the former. We work on simulated systems so that we can focus on the issues that are of most interest to us, involving the kind of architecture sketched above including alarm systems, leaving details of sensory devices and motors till later. When simulations are well designed they can sometimes provide cheaper and faster forms of experimentation, though care is always necessary in extrapolating from simulations.

Many toolkits exist to support such exploration, usually based on a particular architecture or class of architectures (e.g. neural net architectures, or SOAR, or PRS). We wished to investigate diverse and increasingly complex architectures, including coexisting reactive and deliberative sub-architectures, along with self-monitoring and self-modifying capabilities, and including layered perceptual and action subsystems. We also wished to explore varying resource-limits imposed on different components of the architecture, so that, for example, we could compare the effects of speeding up or slowing down planning mechanisms relative to the remaining components of an architecture (e.g. in order to investigate various deliberation management strategies, such as "anytime" planning).

To support this exploration we designed and implemented (in the language Pop-11 (Sloman 1996)) the SIM_AGENT toolkit. It is being used at Birmingham for teaching and research, including research on evolutionary experiments, and also at DERA Malvern for designing simulated agents that could be used in training software. An early version of the toolkit developed jointly with Riccardo Poli, was described at ATAL95 (Sloman & Poli 1996). Since then development has continued in response to comments and suggestions from users (Baxter, Hepplewhite, Logan & Sloman. 1998).

The toolkit supports a collection of interacting agents and inanimate objects, where each agent has an internal architecture involving different sorts of coexisting interacting components, including deliberative and reactive components. Not all agents need have the same architecture.

The key idea is that each component within an agent is connected to other components in that agent via a forward-chaining condition-action rulesystem. Each agent's rulesystem is divided into a collection of different rulesets, where each ruleset is concerned with a specific function, e.g. analysing a type of sensory data, interpreting linguistic messages, creating, checking or executing plans, generating motives, etc. Rulesets can be concurrently active, and may be dynamically switched on and off. They may be assigned different resource limits.

Conditions and actions of rules within an agent can refer to databases in that agent. Thus one form of communication between sub-mechanisms is through the databases in the agent. It is possible for an agent to have some global databases accessed by all components of an agent and others which are used only by specific sub-groups. One agent cannot normally inspect another's databases.

An architecture for an agent class is defined by specifying a collection of rulesets and other mechanisms, along with the types of databases, sensor methods, action methods, communication methods and possibly tracing and debugging methods. It is hoped that users will develop re-usable libraries defining different mechanisms and architectures.

The rulesets are implemented in Poprulebase, a flexible and extendable forward-chaining rule-interpreter. Rulesets can be turned on and off dynamically, modelling one aspect of attention shift, and new ones added, modelling some forms of cognitive development. Although the main conditions and actions use patterns matching database components, some conditions and some actions can invoke sub-mechanisms directly implemented in Pop-11, e.g. low level vision or motor-control mechanisms. Other Poplog languages (e.g. Prolog) or external languages (e.g. C, Fortran) can also be invoked in conditions and actions. For example, a rule condition could in principle interrogate physical sensors and a rule action could send signals to motors. Sockets can run sub-systems on other machines, and unix pipes can communicate with processes on the same machine.

To illustrate the power, a Pop-11 rule action can run the rule interpreter recursively on a specialised rule system.

The rule-based formalism is easily extendable, allowing different sorts of condition-action rules to be defined. For example, one of the extensions designed by Riccardo Poli allows a set of conditions matched against a database to provide a set of input values for a neural net, whose output is a boolean vector which can be used to select a subset of actions to be run. A recent extension was a new class of ADD and DELETE actions for automatically maintaining sets of dependency information between database items, so that if an item is deleted then everything recorded as directly or indirectly depending on it, is also deleted. A Pop-11 condition can be used to perform backward chaining if desired.

The interpreter can be run with various control strategies, including the following options for each active ruleset on each cycle: (a) all runnable rules (those with all conditions satisfied) are run, (b) only the first runnable rule found is run, (c) the set of runnable rule instances is sorted and pruned (using a user-defined procedure) before the actions are run.

When the rule interpreter is applied to a ruleset, it can be allowed to run to completion (e.g. until no more rules have all conditions satisfied, or a "STOP" action is executed.) Alternatively it can be run with a cycle limit N, specifying that it should be suspended after N cycles even if there are still rules with satisfied conditions. Another possibility is to set a timer and halt it after a fixed time interval. Either of these mechanisms can be used to impose resource limits on one ruleset relative to others, within an agent.

The design of the toolkit supports multi-agent scenarios, using a time-sliced scheduler which in each time slice allows each agent to run its sensory methods, its internal rulesets, and, in a second pass at the end of the time slice, its *external* action methods.

The object oriented design uses Pop-11's Objectclass system, which supports multiple inheritance and generic functions. This makes it easy for users to extend the ontology by defining new sub-classes, with their own sensing, acting and internal processing methods, without any editing of the core toolkit code. A default class provides a default set of methods, including the sim_run_agent method used to run each the agent's rulesets, along with various tracing methods.

The object oriented approach allows a Pop-11 graphical library to be connected to the toolkit by re-defining tracing and other methods (e.g. move methods) to invoke graphical procedures. The graphical facilities support not only displays of agent actions but also asynchronous user intervention: e.g. using the mouse to move objects in an agent's environment, or turning tracing and profiling mechanisms on or off while the toolkit is running.

Scenarios implemented so far using the toolkit include a simulated robot using a hybrid modular architecture to propel a boat to follow the walls of an irregular room, evolution of a primitive language for cooperation between a blind and an immobile agent, a user controlled sheepdog and sheep to be penned, two purely reactive "teams" of agents able to move past each other and static obstacles to get to their target locations, a simulated nursemaid looking after troublesome infants while performing a construction task, a distributed minder (Davis 1996), one agent tracking another subject to path constraints in 3-D undulating terrain, and, at DERA Malvern, simulated tank commanders and tank drivers engaging in battle scenarios (Baxter 1996). We expect to continue developing the toolkit and building increasingly sophisticated simulations, moving towards the architecture depicted in Figure 1 and subsequently extended in various ways.

In particular we have plans for improving the self modifying and self monitoring capabilities by replacing the rulesystem, currently a list of rulesets and rulefamilies, with database entries. Thus rule actions can then change the processing architecture.

The toolkit is applicable to a wide range of agent development tasks, including simplified software agents which require only a small subset of beliefs, goals, plans, decisions, reactions to unexpected situations, etc. These might be web search agents, or "believable" entertainment agents whose observed behaviour invites mentalistic description whether or not the descriptions are justified by internal mechanisms, states and processes, e.g. the OZ project at CMU (Bates, Loyall & Reilly 1991). The toolkit could also be used to implement teaching and demonstration libraries, e.g. for students in psychology or the helping professions, where students can manipulate the architectures of simplified human-like agents, to gain a deeper understanding of the multiple ways in which things

can go wrong.

CONCLUSION

Like software engineers, and unlike Dennett and Newell, we assume semantically competent sub-systems, but not rationality. Using this information-level design stance, we have sketched a framework accommodating multi-disciplinary investigation of many types of architecture of varying degrees of sophistication, with varying mixtures of information-processing capability, based on AI, Alife, Biology, Neuroscience, Psychology, Psychiatry, Anthropology, Linguistics and Philosophy. This framework can extend our understanding of both natural and artificial agents. Above all it generates systems of concepts for characterising various types of mentality. Information-based control architectures provide a new framework for analysing, justifying and extending familiar mentalistic concepts.

There is no uniquely "right" architecture. Types of architectures that are relevant, and dimensions of possible variation, are not yet well understood. More exploration and analysis is required, replacing premature (sometimes confrontational) commitment to particular mechanisms and strategies. We need to understand the structure of design space and niche space, and trajectories that are possible within those spaces (Sloman 1994*a*, Sloman 1994*b*, Sloman 1998(forthcoming)). This requires collaborative philosophical analysis, psychological and neurophysiological research, experiments with diverse working models of agents, and evolutionary investigations. Some of this exploration can be based in part on powerful new software tools.

Such work is likely to throw up types of architectures that we would not otherwise think of, which will force us to invent new concepts for describing synthetic minds which are not like our own, and help us understand our own by contrast.

ACKNOWLEDGEMENTS & NOTES

We acknowledge support from the UK Joint Council Initiative, The Renaissance Trust, and DERA Malvern, and much help from students and staff at Birmingham, especially Luc Beaudoin and Ian Wright.

Pointers to code and documentation for our toolkit are at http://www.cs.bham.ac.uk/∼axs/cog_affect/sim_agent.html

Several papers developing these ideas are in the Cognition and Affect Project ftp directory: ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect

References

Bates, J., Loyall, A. B. & Reilly, W. S. (1991), Broad agents, *in* 'Paper presented at AAAI spring symposium on integrated intelligent architectures'. (Available in SIGART BULLETIN, 2(4), Aug. 1991, pp. 38–40).

Baxter, J., Hepplewhite, R., Logan, B. & Sloman., A. (1998), Sim_agent two years on, Technical Report CSRP-98-2, University of Birmingham, School of Computer Science.

Baxter, J. W. (1996), Executing plans in a land battlefield simulation, *in* 'Proceedings of the AAAI Fall symposium on Plan execution: Problems and issues November 1996', AAAI, pp. 15–18.

Davis, D. N. (1996), Reactive and motivational agents: Towards a collective minder, *in* J. Muller, M. Wooldridge & N. Jennings, eds, 'Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages', Springer-Verlag.

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

Goleman, D. (1996), *Emotional Intelligence: Why It Can Matter More than IQ*, Bloomsbury Publishing, London.

McCarthy, J. (1979), Ascribing mental qualities to machines, *in* M. Ringle, ed., 'Philosophical Perspectives in Artificial Intelligence', Humanities Press, Atlantic Highlands, NJ, pp. 161–195. (Also accessible at http://www-formal.stanford.edu/jmc/ascribing/ascribing.html).

McCarthy, J. (1995), Making robots conscious of their mental states, *in* 'AAAI Spring Symposium on Representing Mental States and Mechanisms'. Accessible via http://www-formal.stanford.edu/jmc/consciousness.html.

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

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

Sloman, A. (1978), *The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind*, Harvester Press (and Humanities Press), Hassocks, Sussex.

Sloman, A. (1994*a*), Explorations in design space, *in* 'Proceedings 11th European Conference on AI', Amsterdam.

Sloman, A. (1994*b*), 'Semantics in an intelligent control system', *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering* **349**(1689), 43–58.

Sloman, A. (1996), *Primer of Pop-11*, School of Computer Science, University of Birmingham, School of Computer Science, The University of Birmingham, Birmingham, B15 2TT, UK. (Last revised October 1997. Ftp version available in **ftp://ftp.cs.bham.ac.uk/pub/dist/poplog/**).

Sloman, A. (1998(forthcoming)), What sort of architecture is required for a human-like agent?, *in* M. Wooldridge & A. Rao, eds, 'Foundations of Rational Agency', Kluwer Academic.

Sloman, A. & Poli, R. (1996), Sim_agent: A toolkit for exploring agent designs, *in* M. Wooldridge, J. Mueller & M. Tambe, eds, 'Intelligent Agents Vol II (ATAL-95)', Springer-Verlag, pp. 392–407.

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