Cognition and Affect: Architectures and Tools

Aaron Sloman and Brian Logan
School of Computer Science, University of Birmingham
Birmingham B15 2TT UK
{a.sloman,b.s.logan}@cs.bham.ac.uk

Abstract

Which agent architectures are capable of justifying descriptions in terms of the 'higher level' mental concepts applicable to human beings? We propose a new kind of architecture-based semantics for mentalistic descriptions in which mental concepts (e.g. 'believes', 'desires', 'intends', 'mood', 'emotion', etc.) are grounded in assumptions about information processing architectures, and not merely in concepts based solely on Dennett's 'intentional stance'. These ideas have led to the design of the SIM_AGENT toolkit which has been used to explore a variety of such architectures.

1 Architectures and mentalistic descriptions

McCarthy [1] gives reasons why we shall need to describe intelligent robots in mentalistic terms, and why such a robot will need some degree of self consciousness. He also proposes a notation that we and the robot might use to describe its states. We extend that work by focusing on the underlying 'high level' architectures required to justify ascriptions of mentality. This provides a new kind of architecture-based semantics for mentalistic descriptions. The types of architectures that are relevant are not yet well understood and more exploration and analysis is required.

Different views of a system often reveal different architectures, e.g. a physical architecture, a physiological architecture, and an information processing architecture. There need not be simple correspondences between components of these architectures. Our work deals with an information processing control architecture involved in producing various kinds of internal and external behaviour of the system, including external physical actions and more subtle processes such as motive generation, attention switching, problem solving, information storage, skill acquisition, or even modification of the architecture.

Our main methodological point is that we need further investigation of types of architectures required to support familiar systems of concepts. High-level descriptive concepts applicable to an organism, software agent, or robot, will depend on its architecture. Concepts based entirely on observed behavioural patterns which bear no relationship to the architecture, are likely to be shallow and lacking in predictive and explanatory power. For example, much recent work, based on e.g. [2], adopts a 'parameter based' model of emotion in which emotional states are produced by a distinct component of an agent's architecture, and modify the behavioural responses of the agent. In contrast, we conjecture that some emotional states arise from the interaction of other components of the architecture concerned with motive generation, attention switching, problem solving and so on.

Our aim is to identify high level functional divisions relevant to generating families of mental states and processes. This extends the approach of Dennett, who recommends the 'intentional stance', which restricts mentalistic language to descriptions of whole agents, and presupposes that the agents are largely rational. By contrast, we claim that such talk primarily refers to an 'information level' architecture, close to the requirements specified by software engineers. This uses a design level of description which lies between physical levels (including physical design levels) of description and intentional descriptions that always refer to the whole agent. (The fashionable alternative view of architectures composed only of large and complex collections of very low level components, without any modularity at intermediate levels, if correct, would make a scientific psychology impossible. We suggest that the trade-offs involved in such systems are likely to have prevented their evolution, except in animals with rigidly restricted functionality, e.g. insects.)

The 'holistic' intentional stance includes only what *the whole* (rational) agent believes, desires, intends, etc., whereas 'information level' design descriptions permit reference to various *internal* semantically rich information structures and information processes. This includes sensory buffers, longer term stored associations, generalisations about the environment and the agent, stored information about the local environment, currently active motives, motive generators producing motives under various conditions, mechanisms and rules for detecting and resolving conflicts, learnt automatic responses, planning mechanisms, previously constructed plans or plan schemata, high level control states which can modulate the behaviour of other mechanisms, and many more. In this way we can attempt to explain what (some) mental states are in terms of the information processing and control functions of the architecture.

Our approach is to use whatever evidence is available from brain science, experimental psychology, forms of mental disorder, patterns of development in infancy and decay in old age, evolution and other sources to suggest plausible architectures, which can then be tested by implementing them and running experiments with the implementations, or by performing further empirical research. Initially constraints on theories will be very ill-defined because of paucity of relevant knowledge. However, as more and more constraints come from advances in other fields, more and more tests can be generated to help us choose between alternative hypotheses.

2 Three architectural layers

Like many others, we conjecture that human-like agents need a control architecture with at least three distinct layers which evolved at different times [3]:

- a very old reactive layer, found in various forms in all animals, including insects;
- a more recently evolved deliberative layer, found in varying degrees of sophistication in some other animals (e.g. cats, monkeys);
- an even more recent meta-management (reflective) layer providing self-monitoring, self-evaluation, and self-control, perhaps found only in simple forms in other primates, and perhaps not in other animals.

The layers are not assumed to form a rigidly hierarchical control architecture. Rather the three layers operate concurrently with mutual influences. The reactive mechanisms 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.

We also propose that perceptual and motor systems have evolved layers required for effective interaction with the the three more central layers. 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.

To illustrate our claims about architecture-based concepts, we suggest that the different layers account for different sorts of emotional states, only some of which are shared with other animals [5].

- The reactive layer produces rapid automatically stimulated emotional states (like being startled, terrified, sexually excited).
- The deliberative layer supports cognitively rich emotional states linked to current desires plans and beliefs (like being anxious, apprehensive, relieved, pleasantly surprised).
- Certain emotional states (like feeling humiliated, infatuated, guilty, or full of excited anticipation) involve reduced ability to focus attention on urgent or important tasks because of processes interrupting and diverting deliberative mechanisms, sometimes conflicting with decisions in the meta-management layer.

It may be useful to distinguish 'routine' reactive mechanisms from a 'global alarm' mechanism, such as seems to be implemented in the limbic system in mammals, and plays a role in some (especially primeval) emotions.

3 The SIM_AGENT toolkit

Since we still have a lot to learn about possible types of agent architectures and their properties, and since the properties of complex systems cannot all be determined in a purely theoretical fashion, e.g. by logical and mathematical analysis, there is a need for a great deal more explanation of various types of architectures, both in physical robots and in simulated systems. Provided that the latter are well designed they can sometimes provide cheaper and faster ways of exploring issues relevant to physically implemented systems, though care is always necessary in extrapolating from simulations to "the real thing".

Many toolkits exist to support such exploration. However, many of them are committed to a particular architecture of class of architectures (e.g. behaviour-based architectures, or SOAR, or PRS). In order to investigate a range of increasingly complex and diverse architectures including architectures combining coexisting reactive and deliberative sub-architectures, along with self monitoring capabilities, we have designed and implemented the SIM_AGENT toolkit, which is being used at Birmingham for teaching and research, including research on evolutionary experiments, and at DERA Malvern for designing simulated agents that could be used for training software. [4] We expect to continue developing the toolkit and building increasingly sophisticated simulations.

Acknowledgements and Notes

Our work has been supported by the UK Joint Council Initiative and DERA Malvern, among others. A number of papers developing these ideas, together with the current version of the SIM_AGENT toolkit can be found at:

http://www.cs.bham.ac.uk/~axs/cogaff.html

References

[1] J. McCarthy. Ascribing mental qualities to machines. In M. Ringle, editor, *Philosophical Perspectives in Artificial Intelligence*, pages 161–195. Humanities Press, Atlantic Highlands, NJ, 1979. http://www-formal.stanford.edu/jmc/ascribing/ascribing.html.

[2] A. Ortony, G.L. Clore, and A. Collins. *The Cognitive Structure of the Emotions*. Cambridge University Press, New York, 1988.

[3] A. Sloman. What sort of architecture is required for a human-like agent? In Michael Wooldridge and Anand Rao, editors, *Foundations of Rational Agency*, pages 35–52. Kluwer Academic, Dordrecht, 1999. http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#21.

[4] A. Sloman and R. Poli. Sim_agent: A toolkit for exploring agent designs. In M. Wooldridge, J. Mueller, and M. Tambe, editors, *Intelligent Agents Vol II (ATAL-95)*, pages 392–407. Springer-Verlag, 1996.

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