COMPUTATIONAL MODELLING OF MOTIVE-MANAGEMENT PROCESSES

Aaron Sloman, Luc Beaudoin and Ian Wright

School of Computer Science & Cognitive Science Research Centre

The University of Birmingham.

(In collaboration with Glyn Humphreys, Tim Read, Edmund Shing, Christian Paterson).

Introduction

A major task for cognitive science is to design architectures for intelligent agents capable of accounting for the diverse phenomena studied by psychologists and encountered in ordinary life. The explanations must involve the ideas of control and information: in particular control by *information-rich* states (Sloman 1994b). A collection of design schemata should accommodate phenomena common to diverse individuals and be capable of further specification to account for individual differences and abnormalities. Ideally they should also accommodate evolutionary and cross-species studies. This grand long-term task generates many smaller-scale sub-tasks. The Cognition and Affect project at Birmingham is concerned with "global" design requirements for agents that cope simultaneously with coexisting but possibly unrelated goals, desires, preferences, intentions, and other kinds of motivators, all at different stages of processing, including generation of emotions. Our work builds on and extends seminal ideas of H.A.Simon (1967), and other proponents of information-processing models of mind.

There are many approaches to such problems: evolutionary, neurobiological, experimental, philosophical and "design-based" (Sloman 1993a). We adopt a design-based exploration of sets of requirements and designs matching them. This is part of a longer term study of mappings between "niche-space" (the space of sets of requirements) and "design-space" (Sloman 1994a). The requirements include coping with the fact that new sensory information, and new motivators can turn up at any time, causing interrupts, conflicts, or new opportunities. At a physical level, incompatible actions may be required. At higher levels, goals and preferences may be incompatible and, because processing resources are limited, there may not be time for all urgent or important processes to be completed.

Design-based work relating designs to requirements and constraints should lead to a deeper understanding of many types of human mental states and processes and provide a richer and more precise set of concepts for talking about them, for doing cross-species comparisons, for explaining evolutionary pressures, for guiding neuro-biological research and for understanding malfunctions of many kinds.

More detailed objectives: An application of the design stance

The project has the following more specific objectives:

1. To identify high level functional requirements for architectures for human-like intelligent agents, especially requirements concerned with real time processing of asynchronously generated motivators and percepts, and control of attention in resource-limited agents. (More generally: to identify interesting sets of requirements and constraints, including different kinds of resource limits and their implications.)

2. To explore possible designs capable of fulfilling those requirements, and the trade-offs between them, and to investigate consequences of various design options. (This may help us understand differences between organisms.)

3. To implement a sequence of designs capable of fulfilling increasing subsets of those requirements, in order to explore and demonstrate their implications. Design ideas are extended and tested through computational modelling. (Often the very process of design and implementation reveals flaws and gaps in a theory, before any programs are actually run.)

4. To use the generative power of proposed mechanisms as a basis for building a conceptual framework for describing high level mental states, including affective states involving control (or loss of control) of attention. This provides a new basis for philosophical analysis of mental concepts.

5. To design effective interfaces for demonstrating the key features of the models, and to enable them to be used for teaching or training.

Broad and shallow architectures

Among the many possible approaches to exploration of designs, our group concentrates (for the time being) on top-down design of "broad" architectures, i.e. architectures combining many kinds of functionality apparently found in people. This means that most components in the architecture are drastically over-simplified. The architecture is "broad and shallow" (like the work of Bates et al. 1991). Later work will incorporate more complete and realistic components, moving towards "broad and deep" architectures.

One of the main design constraints is that processing resources are limited, at least for high level cognitively-based processes such as deliberation, planning, and problem-solving. Low-level sensory processes, e.g. in vision, appear to use vast computing resources. By contrast, deliberative processes in which alternatives are explicitly represented and evaluated, and new structures created (plans, sentences, arguments), seem to have limited parallelism. You can walk and talk and enjoy the scenery in parallel, but not say five different poems to yourself in parallel (why not?). Resource limits imply a need for resource allocation, i.e. selection between options. We construe *attention* as being concerned with making selections at all levels: between *what* information to process, *how* to process it, *which* new goals to consider, *how* to consider them, *which* to adopt, *how* to attain them, and so on. The project has focused on these "management" and "meta-management" processes and their interaction with multiple sources of motivation and potentially disruptive new information. (See Figures 1 and 2, but treat them merely as imprecise suggestive sketches.)

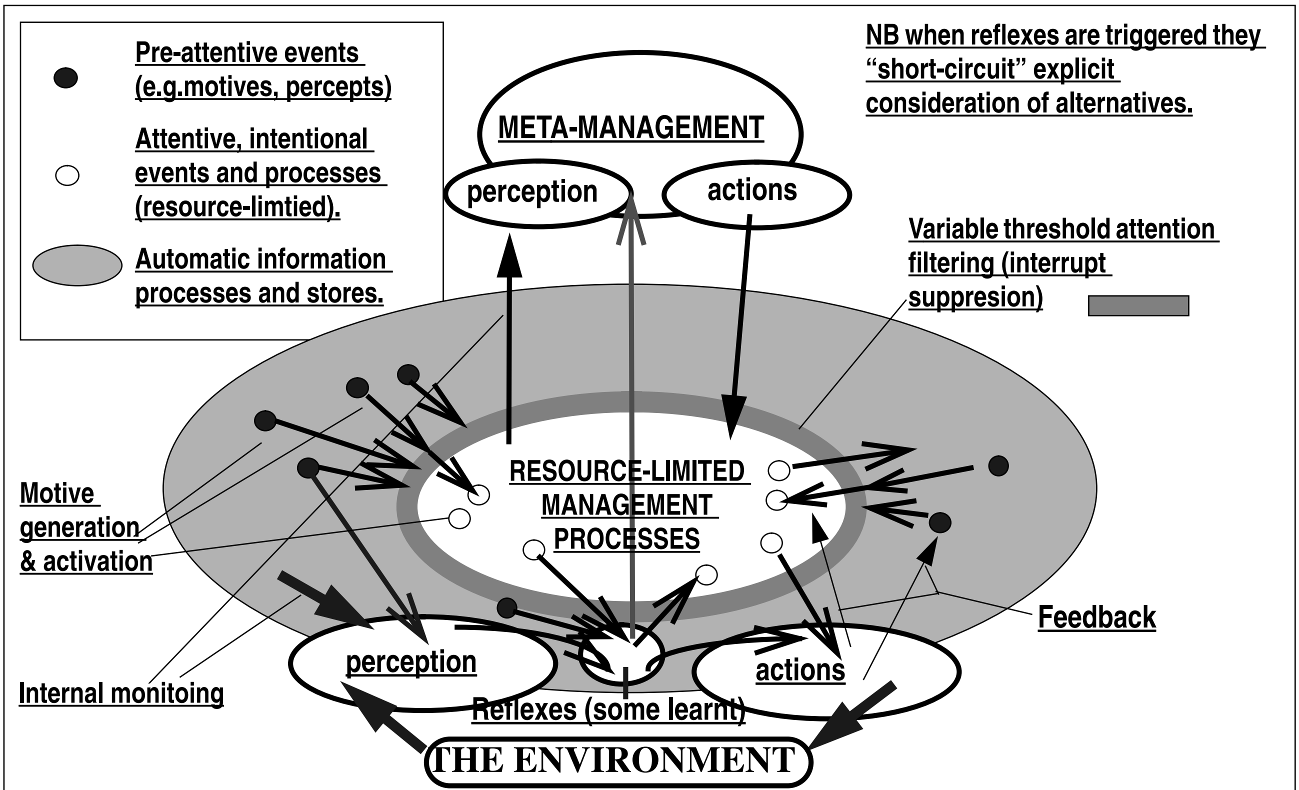


FIG 1: TOWARDS AN ARCHITECTURE FOR AN INTELLIGENT AGENT

(In a deeper architecture, perception would be multi-layered)

Some processes involving motivators (desires, goals, intentions, etc.)

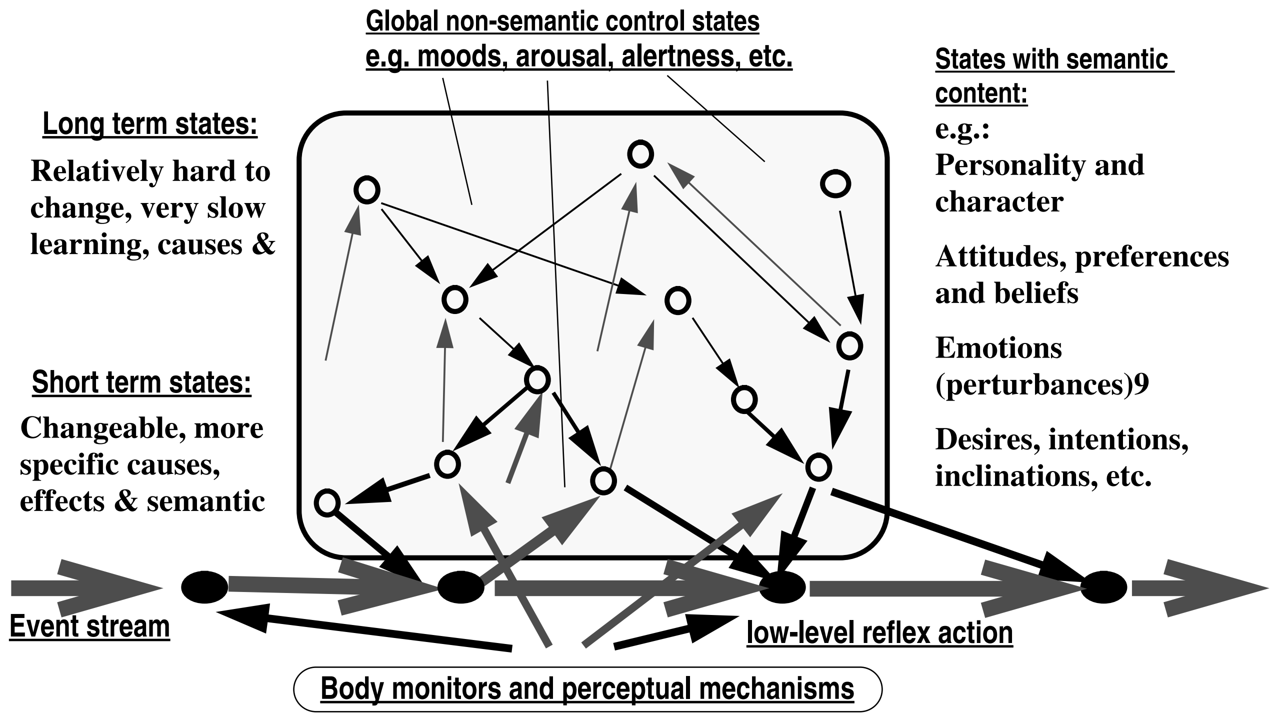
The following are among the processes to be explained:

- **Motivator generation and (re-)activation**. (Includes using pre-attentive "generactivators".)
- **Suppression or 'filtering' mechanisms**: to prevent management processes from being diverted by active motivators when diversion would have disastrous effects. Filtering new motives is one means of controlling attention. The filtering decisions must not themselves consume significant resources: so they use heuristics. (See the architecture diagram (Fig 1)) Heuristics can lead to "bad" decisions about what should get through and interrupt
- **Management of motivators** (including the following:)
  - **Assessing motivators.** Motivators and their possible expansions are assessed along dimensions of importance (evaluation of achievement or failure and their consequences), likelihood of satisfaction, cost of satisfaction, urgency, etc. These dimensions of evaluation are distinguished by the different effects they have on processing.
  - **Deciding**: deciding whether to adopt the motivator, i.e. form an intention.
  - **Scheduling**: deciding when or under which conditions to execute a motivator.
  - **Expansion**: deciding how to execute a motivator (expanding goals into plans, including partial/ schematic plans).
  - **Predicting the effects** of (hypothetical) decisions, or plan execution.
  - **Assigning "intensity" measure:** i.e. a measure of a tendency of a goal to gain and retain control once it has been adopted.
  - **Detecting conflicts between motivators.**
  - **Detecting mutual support between motivators.**
  - **Setting thresholds for the management interrupt filter.**

- **Termination of motivators.** Contrast explicit termination (e.g. due to goal satisfaction) and decay, e.g. due to the source no longer being present, or other pre-attentional processes.

- **Meta-management:** processes that (recursively) control management or meta-management processes. E.g. deciding whether to continue planning how to achieve X, deciding when to think about Y, or deciding to abandon deliberation about Z. Effective meta-management depends on a self-model and self-monitoring. (This can vary in accuracy and completeness.)
- **Execution of plans**, with or without high level management. This includes run-time monitoring and planning.
- **Learning** to improve these processes (e.g. improving planning or filtering strategies). Some learning *improves* performance. Some *extends* it.
- **Extending the architecture.** E.g. this could involve learning new skills, or developing new links between sub-processes, including new "cognitive reflexes".

Other processes include global "switches" or "modulators" changing processing either quantitatively or qualitatively. Many of these will have no semantic content (e.g. mood changes, arousal changes), though they may result from semantic processing. See Fig 2. There is enormous scope for individual variability in the architecture and its behaviour.



Heavy barbed horizontal arrows and black circles represent events ordered in time.

Non-horizontal arrows represent causes of differing strengths, differing time-scales (learning), some deterministic, some probabilistic.

White circles represent control states, some long term, some short term, some specific and goal-directed, e.g. desires.

Some control states (e.g. attitudes, desires) have semantic content and generate specific goals and decisions, while others are more global and diffuse (the shaded area) and merely modulate processing, e.g. speed, risk-taking, alertness, amount of deliberation prior to action, optimism, etc.

Most are subject to both top-down (system-driven, knowledge-driven) and bottom-up (event-driven, sensor-driven) influences

FIG 2: HIERARCHIES OF (DISPOSITIONAL) CONTROL

Motivator structure

Motivators (desires, inclinations, goals, etc.) produced by pre-attentive or attentive processes generally include the following components, though they may have other specific features also. Some of these will vary over time.

(1) Semantic content: a proposition, **P**, denoting a possible state of affairs, which may be true or false
(2) A motivational attitude to **P,** e.g. “make true”, “keep true”, “make false”, etc.
(3) A rationale, if the motivator arose from explicit reasoning.
(4) An indication of the current belief about **P**’s status, e.g. true, false, nearly true, probable, unlikely etc.
(5) An “importance value” (e.g. “neutral”, “low”, “medium”, “high”, “unknown”), importance may be *intrinsic*, or based on assessment of *consequences* of (doing and not doing).
(6) An “urgency descriptor” (possibly a time/cost function)
(7) A heuristically computed “insistence value”, determining interrupt capabilities. Should correspond to estimated importance and urgency.
(8) Intensity -- which influences likelihood of (continuing) being acted on, as against other motivators.
(9) Possibly a plan or set of plans for achieving the motivator
(10) A commitment status (e.g. “adopted”, “rejected”, “undecided”)
(11) A dynamic state (e.g. “being considered”, “consideration deferred till...”, “nearing completion”, etc.)
(12) Management information, e.g. the state of current relevant management and meta-management processes.

In most animals motivators are probably simpler (and in current robots). There may be individual differences among humans too, including processing capabilities. Exploring “design space” will show what is possible. (Sloman 1994a).

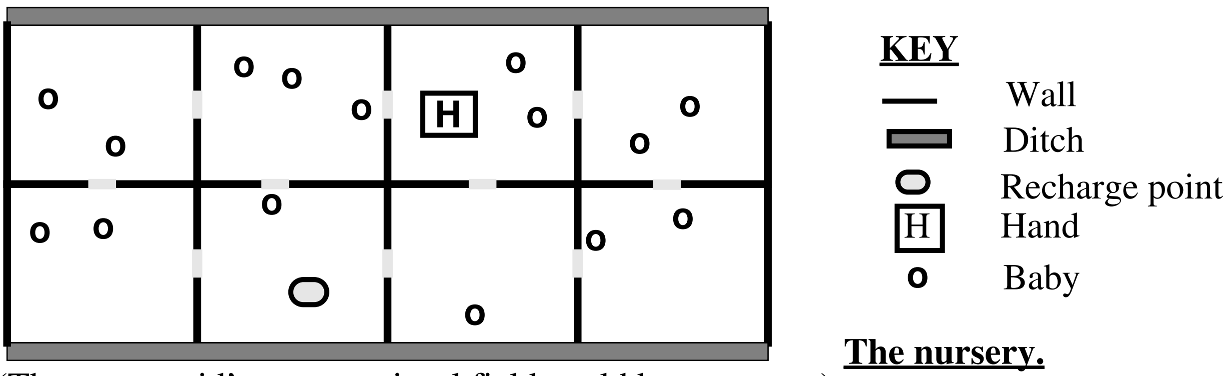
What has this to do with emotions?

A key conjecture is that certain states (“perturbances”) involving partial loss of control of attention emerge as a by-product of the activity of motivator-processing architectures designed (by evolution or, one day, synthetically) to meet the requirements of intelligent but resource-limited agents in an environment as complex as ours. This partial loss of control of attention, disturbing management and meta-management processes, due in part to imperfect interrupt filtering heuristics, seems to be a central characteristic of many human emotional states, e.g. grief, jealousy, infatuation, excited anticipation. The neural basis, and the more “superficial” aspects of emotional states, e.g. facial expression, sweating, muscular tension, production of tears, and other physiological changes are (for now) left for others to investigate.

There are such diverse definitions of “emotion” that arguing which is correct is a waste of time (Read 1993). We find it more fruitful to use detailed descriptions of the states and processes we wish to account for. At present we are mainly concerned with high level interactions between beliefs, motivators (desires, intentions, preferences, etc.), plans, and attentional processes, and ways in which they can function under stress, or possibly go wrong, e.g.poor management of coexisting processes, problems due to excessive “busyness”, etc. Later work will accommodate innate or learnt short-circuits, e.g. cognitive reflexes, and more global (semantics free) control states. The “global interrupt” signals postulated by Oatley and Johnson-Laird when plans meet obstacles etc. would be a special case. A good theory of the architecture of an agent should generate more comprehensive and precise concepts than those corresponding to current usage of words like “emotion”, “mood”, “desire”, “attitude”, “belief”, “personality”, “memory”, “perception” etc.

The nursemaid (minder) scenario

In order to develop and test these ideas we use a simulated domain in which an agent has to perform multiple tasks under time pressure, with incomplete or uncertain information. We chose a “toy” 2-D domain presenting problems we



The nursery.

(The nursemaid’s current visual field could be any room)

wished to address while avoiding others. The domain had to be simple (initially) in order to make the problems tractable without first solving all the problems of AI, including 3-D vision, motor control, and naive physics. We chose a “toy” 2-D domain presenting problems we wished to address while avoiding others. It supports exploration of “Broad and Shallow” architectures: broad because of diverse functions, shallow because individual components are (at first) simplified. Progressive deepening will come later. In the scenario a “nursemaid” with a movable eye and a hand, has the task of looking after a collection of mobile “babies” in a 2-D world with various dangers and opportunities. A first draft architecture for the nursemaid’s “mind” has been developed and partially implemented by Luc Beaudoin and described in his PhD thesis. Ian Wright is re-implementing it and extending it to support internal self-monitoring of the management and meta-management processes. This will enrich the meta-management capabilities. (The implementation uses Pop-11 and X in the Poplog environment running on a Sparcstation 10/30. Not everything sketched here has already been implemented.)

Future plans

Our work will grow in several different directions, combining the “design-based” approach with clinical and other approaches, in order to explore the structure of niche-space, and the space of possible designs capable of matching various sets of requirements (more or less satisfactorily). Cross-disciplinary workshops will enable us to share ideas.

Elaboration of the nursemaid domain could include extending the topography, complicating the variety of processes, giving the nursemaid a more complex “body”, adding a social dimension, adding more forms of learning, and so on.

Theoretical work should include enrich the architecture in various ways, e.g. to include self-monitoring and a more complete “self model”, with more forms of learning, including cognitive reflexes, “compiled plans”, mechanisms for pain and pleasure, etc.

Potential engineering applications are concerned with the design of resource-limited intelligent systems that have to operate in demanding time-critical and partly unpredictable domains, e.g. command and control systems and decision-support systems.

Computer implementations could provide the basis for tutorial programs for teaching students of psychology, psychotherapy, etc. about some of the processes they may need to think about, and perhaps training managers, etc.

The architecture should provide a systematic framework for generating new, more precise, concepts describing states and processes that can occur (e.g. emotions, moods, attitudes, personality, etc.), much as the theory of the architecture of matter systematically generates concepts of chemical elements and compounds, and their behaviour.

Short Bibliography

Bates, J., Loyall, A. B., & Reilly, W. S. (1991). Broad agents. *Proceedings AAAI spring symposium on integrated intelligent architectures*. Stanford, CA: Reprinted in *Sigart Bulletin,* 2(4), Aug. 1991, pp. 38-40.).

Beaudoin, LP., (submitted 1994) *A design-based study of autonomous agents*, PhD thesis, School of Computer Science The University of Birmingham.

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

Johnson-Laird, P. N.(1988) *The Computer and the Mind: An Introduction to Cognitive Science*, Fontana

Oatley, K. and P.N. Johnson-Laird (1987), Towards a cognitive theory of emotions, *Cognition and Emotion*, 1: 29-50.

Read, T. & Sloman A, (1993) ‘The terminological pitfalls of studying emotion’ in *Proceedings WAUME93: Workshop on Architectures Underlying Motivation and Emotion,* (see below)

Read, T. & Sloman A.(eds) *Proceedings WAUME93: Workshop on Architectures Underlying Motivation and Emotion,* Aug. 11-12 1993} Cognitive Science Research Papers, University of Birmingham. (Mainly abstracts)

Simon, H.A.(1967) ‘Motivational and Emotional Controls of Cognition’, reprinted in *Models of Thought*, Yale University Press, (1979) 29--38.

Sloman, A, (1978) *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind*, Harvester Press, and Humanities Press. (Out of print, but still in some libraries).

Sloman, A and Monica Croucher, (1981) ‘Why robots will have emotions’, in *Proceedings 7th International Joint Conference on Artificial Intelligence,* Vancouver, 1981.

Sloman, A. (1987) ‘Motives Mechanisms Emotions’ in *Cognition and Emotion*, 1,3, pp.217-234,
reprinted in M.A. Boden (ed) *The Philosophy of Artificial Intelligence,* Oxford Readings in Philosophy Series, Oxford University Press, pp 231-247 1990.

Sloman, A (1992) ‘Prolegomena to a theory of communication and affect’ in Ortony, A., Slack, J., and Stock, O. (Eds.) *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues.* Heidelberg, Germany: Springer, 1992, pp 229-260. (Also available as Cognitive Science Research Paper, The University of Birmingham.)

Sloman, A, (1993a) Prospects for AI as the general science of intelligence’ in A.Sloman, D.Hogg, G.Humphreys, D. Partridge, A. Ramsay (eds) *Prospects for Artificial Intelligence*, IOS Press, Amsterdam, pp 1-10,

Sloman, A, (1993b) ‘The mind as a control system’, in *Philosophy and the Cognitive Sciences,* (eds) C. Hookway and D. Peterson, Cambridge University Press, pp 69-110 (Supplement to Philosophy)

Sloman, A, (1994a), Explorations in design space in *Proc ECAI94, 11th European Conference on Artificial Intelligence*, ed A.G.Cohn, John Wiley, pp 578-582.

Sloman, A (1994b), Semantics in an intelligent control system, in *Proc. British Academy and Royal Society Conference: Artificial Intelligence and The Mind: New Breakthroughs Or Dead Ends?* (April 1994) to appear in Philosophical Transactions of the Royal Society, 1994.

Most of our papers are accessible from our FTP site via WWW: **http://www.cs.bham.ac.uk/~axs/cogaff.html**

or at ftp://ftp.cs.bham.ac.uk/pub/dist/papers/cog_affect. Enquiries to A.Sloman@cs.bham.ac.uk

Many thanks to Dave Moffat for useful comments on presentation.