DFKI Saarbrücken 6th Feb 1997

Architectural Requirements for Autonomous Human-like Agents

Aaron Sloman
School of Computer Science
The University of Birmingham, UK

Email: A.Sloman@cs.bham.ac.uk
www.cs.bham.ac.uk/ ~axs

TALKING ABOUT MENTAL STATES

McCarthy 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, and he has made suggestions regarding the notation that we and the robot might use to describe its states.

This talk extends that work by focusing on the underlying "high level" architectures required to justify ascriptions of mentality.

Which concepts are applicable to a system will depend on the architecture of that system.

An architecture provides a basis for a family of inter-related concepts namely the concepts that describe the states and processes able to occur in the architecture.

EXAMPLE: SELF-CONTROL AND EMOTIONS

We talk about humans sometimes losing control of themselves, for instance in certain emotional states. This presupposes the possibility of switching between being in control and losing self control, which in turn depends on the existence of an architecture that supports certain kinds of self monitoring, self evaluation, and self modification.

For systems lacking the architectural underpinnings, certain descriptions of mental states and processes (e.g. "emotional", "restrained", "resisting temptation") may be inapplicable.

Whether other animals have architectures that can support these descriptions is not clear. Neither is it clear what sorts of architectures in software agents will make such states and processes possible. We have some tentative suggestions outlined below.

A comparison: the architecture of matter

Within the framework of the atomic theory of matter it became possible to see which previous concepts of "kinds of stuff" were suited to describing the physical world and which ones needed to be refined or rejected.

The new architecture also revealed the need for a host of concepts for kinds of physical matter that had not previously been thought of, e.g. elements whose possibility was first revealed by the periodic table.

MENTALISTIC CONCEPTS APPLICABLE TO ARTIFICIAL AGENTS

It is often convenient to describe a machine as:

"choosing", "exploring", "deciding", "inferring", "believing",

and one day perhaps as:

"wanting", "preferring", "enjoying", "disliking", "frightened", "angry", "relieved", "delighted", etc.

The states and processes referred to are *intentional*, **since they have semantic contents.**

If applying such mentalistic concepts to people assumes a certain sort of high level information processing architecture, then similar architectural requirements will need to be satisfied by artificial agents.

Otherwise applying mentalistic terms to them is misleading, like the over-enthusiastic use of words like "goal" and "plan" in some AI publications, criticised by McDermott in 1981.

All this assumes that purely behavioural definitions of mentalistic concepts (in terms of relationships between externally observable inputs and outputs) cannot adequately define these concepts.

WHY USE MENTALISTIC LANGUAGE?

WE NEED MENTALISTIC DESCRIPTIONS:

(a) because of marketing requirements,

(b) because such descriptions will be irresistible

(c) because no other vocabulary will be as useful for describing, explaining, predicting capabilities and behaviour.

Instead of trying to avoid the use of mentalistic language, we need a disciplined approach to its use.

This can come by basing mentalistic concepts on architectural concepts: i.e. we use the 'design stance'.

This differs from the approach of Dennett who recommends the "intentional stance" in describing sophisticated robots, as well as human beings. This stance presupposes that the agents being described are rational.

It also differs from the approach of Newell's "knowledge level" which also presupposes rationality.

THE "INFORMATION LEVEL" DESIGN STANCE
Mentality is concerned with an "information level" architecture, close to the requirements specified by software engineers.

EXAMPLE: CARELESSNESS?

Describing X as "working carelessly" implies

(a) X had certain capabilities relevant to the task

(b) X had the ability to check and detect the need to deploy them

(c) the actual task required them to be deployed

(d) something was lacking in the exercise of these capabilities on this occasion so that some undesirable consequence ensued or nearly ensued.

X's carelessness could have several forms:

- **X forgets the relevance of some of the checks (a memory failure),**
- **X does not focus attention on the data that could indicate the need for remedial action (an attention failure),**
- **X uses some shortcut algorithm that works in some situations and was wrongly judged appropriate here (a selection error),**
- **X does not process the data in sufficient depth because of a misjudgement about the depth required (a strategy failure),**
- **X failed to set up the conditions (e.g. turning on a monitor) that would enable the problem to catch his attention (a management failure).**

The presuppositions for "working carefully" are similar.

Something incapable of being careless cannot be careful.

ARCHITECTURAL LAYERS

A TASK FOR AGENT THEORISTS:

to devise a theory of possible types of architectures and use the architectures 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.

CONJECTURE:

Human-like agents need an architecture with at least three layers.

- **A very old reactive layer, found in various forms in all animals, including insects).**
- **More recently evolved deliberative layer, found in varying degrees of sophistication other animals.**
- **An even more recent meta-management (reflective) layer providing self-monitoring and self-control, perhaps found in simple forms only in other primates. (Probably not in very young children?)**

DIFFERENT LAYERS EXPLAIN DIFFERENT SORTS OF MENTAL STATES AND PROCESSES

For example:

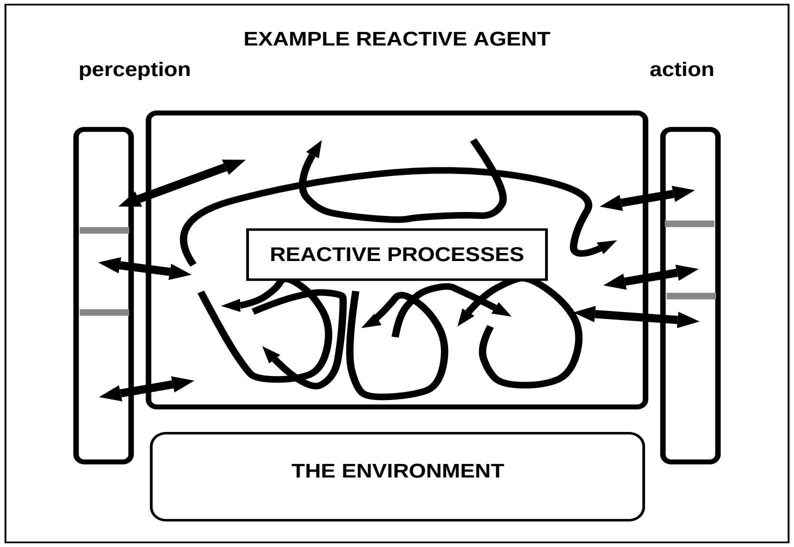
(1) emotional states (like being startled, terrified, sexually stimulated) based on the old reactive layer shared with many other animals,

(2) emotional states (like being anxious, apprehensive, relieved, pleasantly surprised) which depend on the existence of the deliberative layer, in which plans can be created and executed,

(3) emotional states (like feeling humiliated, infatuated, guilty, or full of excited anticipation) in which attempts to focus attention on urgent or important tasks can be difficult or impossible, because of processes involving the meta-management layer.

Within this framework we can dispose of a considerable amount of argumentation at cross-purposes, because people are talking about different sorts of things without a theoretical framework in which to discuss the differences.

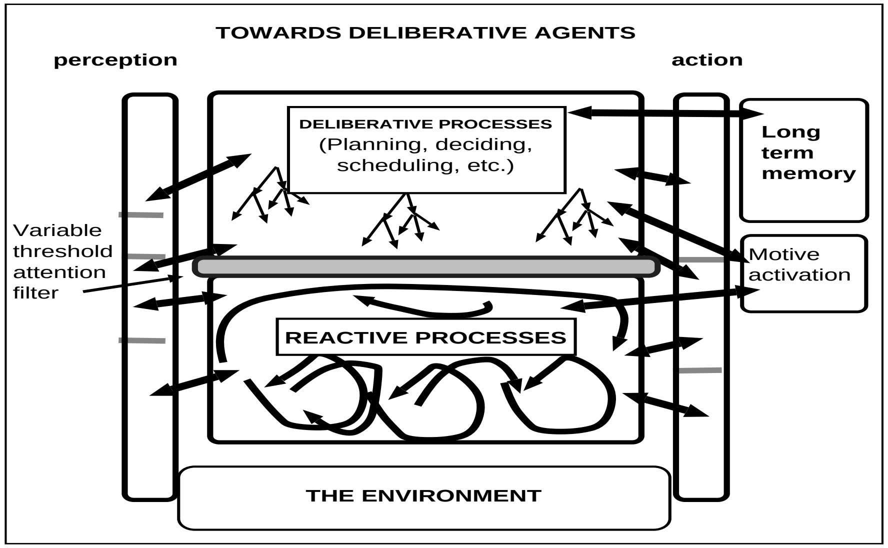
REACTIVE AGENTS



IN A REACTIVE AGENT:

- **Mechanisms and space are dedicated to specific tasks**
- **There is no construction of new plans or structural descriptions**
- **There is no explicit evaluation of alternative structures**
- **Conflicts may be handled by vector addition or winner-takes-all nets.**
- **Parallelism and dedicated hardware give speed**
- **Some learning is possible: e.g. tunable control loops, change of weights by reinforcement learning**
- **The agent can survive even if it has only genetically determined behaviours**
- **Difficulties arise if the environment requires new plan structures.**
- **This may not matter if individuals are cheap and expendable (insects?).**

REACTIVE AND DELIBERATIVE LAYERS



IN A DELIBERATIVE MECHANISM:

- New options are constructed and evaluated
- Mechanisms and space are reused serially
- Learnt skills can be transferred to the reactive layer
- Sensory and action mechanisms may produce or accept more abstract descriptions
- Parallelism is much reduced (for various reasons):
  - Learning requires limited complexity
  - Access to associative memory
  - Integrated control
- A fast-changing environment can cause too many interrupts, frequent re-directions.
- Filtering via dynamically varying thresholds helps but does not solve all problems.

SELF-MONITORING (META-MANAGEMENT)

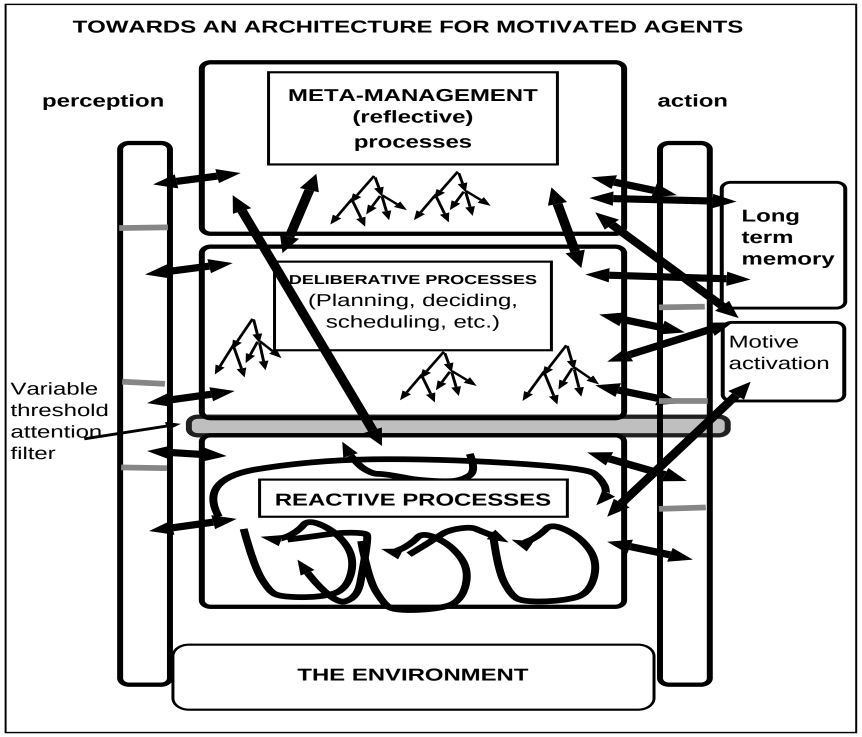
Deliberative mechanisms with evolutionarily determined strategies may be too rigid.

Internal monitoring mechanisms may help to overcome this if they

- Improve the allocation of scarce deliberative resources
- Record events, problems, decisions taken by the deliberative mechanism,
- Detect patterns, such as that certain deliberative strategies work well only in certain conditions,
- Allow exploration of new internal strategies, concepts, evaluation procedures, allowing discovery of new features, generalisations, categorisations,
- Allow diagnosis of injuries and illness by describing internal symptoms to experts,
- Evaluate high level strategies, relative to high level long term generic objectives, or standards.
- Communicate more effectively with others, e.g. by using viewpoint-centred appearances to help direct attention, or using drawings to communicate about how things look.

Meta-meta-management may not be needed if meta-management mechanisms are recursive!

TOWARDS MULTI-LAYERED AUTONOMOUS (REFLECTIVE) AGENTS



"META-MANAGEMENT" PROCESSES MIGHT:

- **Reduce frequency of failure in tasks**
- **Not allow one goal to interfere with other goals**
- **Prevent wasting time on problems that turn out not to be solvable**
- **Reject a slow and resource-consuming strategy if a faster or more elegant one is available**
- **Detecting possibilities for structure sharing among actions.**

ARCHITECTURE AND EMOTION

Different architectural layers support different sorts of emotions:

The reactive layer supports: • being startled
- **being disgusted by horrible sights and smells**
- **being terrified by large fast-approaching objects?**
- **sexual arousal? Aesthetic arousal ?**

The deliberative layer supports: • being frustrated by failure
- **being relieved at avoiding danger**
- **being anxious about things going wrong**
- **being pleasantly surprised by success**

The self monitoring meta-management layer, supports:
- **having and losing control of thoughts and attention:**

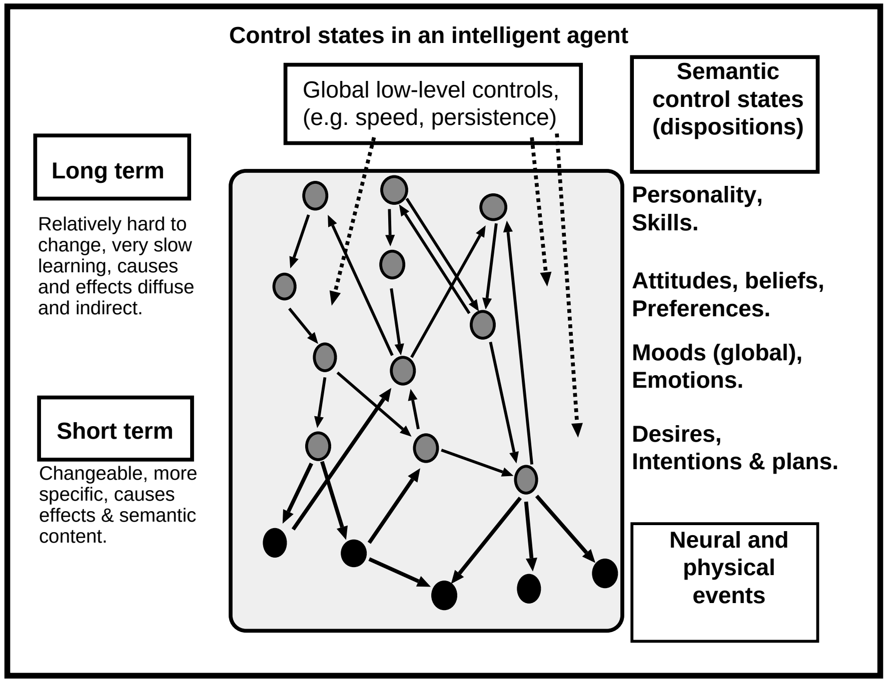
  Feeling ashamed of oneself
  Feeling humiliated
  Aspects of grief, anger, excited anticipation, pride, and many more HUMAN emotions.

NOT EVERYTHING SUPPORTED BY A MECHANISM IS PART OF ITS FUNCTION: MULTI-PROCESSING OPERATING SYSTEMS SUPPORT THRASHING!

SOME FUNCTIONAL MECHANISMS HAVE DYSFUNCTIONAL CONSEQUENCES.

TYPES OF CONTROL STATES

Besides emotions there are personality, attitudes, moods, desires, wishes, intentions, etc.



Control states of varying scope and duration

The "higher" states are:

- **Harder to change**
- **More long lasting**
- **Subject to more influences**
- **More general in their effects**
- **More indirect in their effects**
- **More likely to be genetically determined(??)**

THERE IS NO UNIQUE ARCHITECTURE

Many architectures are needed for different organisms or artificial agents.

Even humans differ from one another: children, adolescents, adults and senile adults.

Naturally occurring alien intelligences and artificial human-like agents may turn out to have architectures that are not exactly like those of normal adult humans.

Different architectures support different classes of mental states.

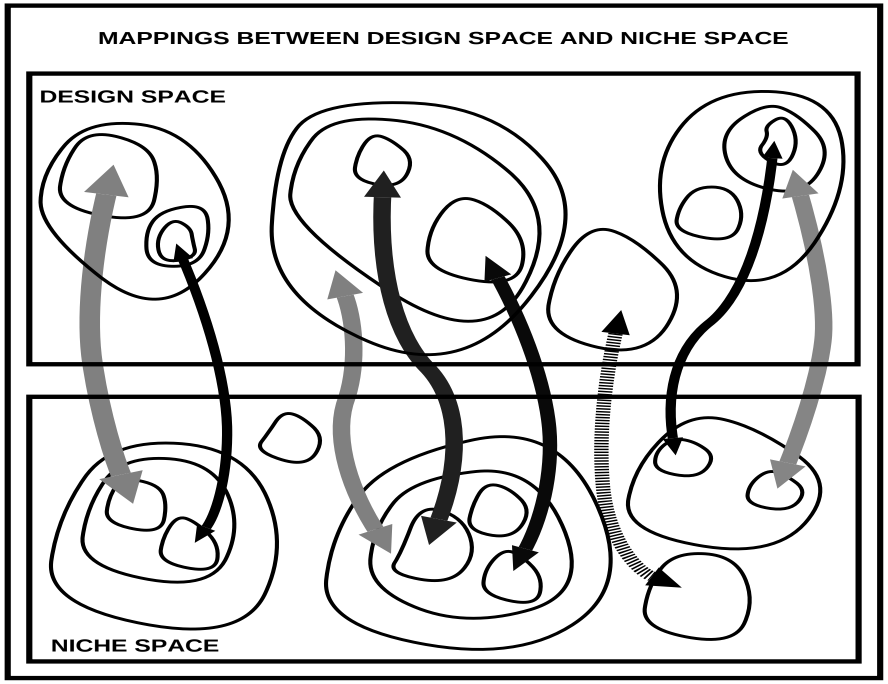
Designers of synthetic agents need to be aware of the evolutionary pressures behind human architectures. Some artificial agents may need similar architectures.

There may be some unanticipated consequences of these design features (Sloman and Croucher IJCAI 1981).

Analysing these possibilities is hard.

So we need to explore relationships between "niche space" and "design space".

DESIGN SPACE and NICHE SPACE



- A niche is a set of requirements
- A design is a set of specifications
- Mappings are not unique: trade-offs everywhere
- Designs need no designer, requirements no requirer.

DYNAMICS:
Which trajectories are possible –

- Within an agent (development, learning)?
- Across generations (evolution, ALIFE)?

The "Turing test" defines a tiny niche region ....
of relatively little interest, except as a technical challenge.

MORE ON THE INFORMATION LEVEL

Information level analysis presupposes that there are various information rich internal structures within the architecture.

These need not be physically demarcated: they could be interacting structures in a virtual machine.

The functional rules of such structures and substates are determined by:

(a) where the information comes from,

(b) how it is stored,

(c) how it is processed or transformed before, during and after storage,

(d) whether it is preserved for a short or long time,

(e) how it can be accessed,

(f) which other components can access it,

(g) what they can do with the information,

(h) whether it actively generates new processes

and so on.

Notions of belief, imagining, reasoning, questioning, pondering, desiring, deciding, intending, having a mood, having an attitude, being emotional, etc. all presuppose diverse information stores with diverse syntactic forms, diverse mechanisms for operating on them, diverse contents and functional roles within the architecture.

WHY RATIONALITY IS NOT PRESUPPOSED

These mental states do not presuppose rationality because many interactions between the components can produce irrational decisions or actions.

For instance irrational impulses can be a product of an information processing architecture part of which is highly reactive.

Moreover, there are many forms of partial breakdown, corruption of memory stores, distraction while executing strategies, etc.

ALIEN INTELLIGENCES

If some architectures are too different from our own we may need new sorts of concepts for describing their states and explaining their behaviour.

This includes some already found in nature.

Can a goldfish long for its mother, and if not why not?

Can we talk about what a fly sees?

In some artificial agents, our normal modes of description may not be appropriate.

For those we'll need to develop new systems of concepts and explanatory principles.

ARE THE PROBLEMS TOO HARD?

ANSWER: Use the 'Broad and Shallow' Approach (Bates)

- **Explore many interacting components**
- **Most (all) components initially have shallow implementations (e.g. perception, planning)**
- **Progressive deepening**
- **Explore different starting points**
- **Initially: Broad and Very Very Shallow (BVVS)**
- **Combine as many approaches as possible, learning from others**
- **Develop shared libraries**

Also try to develop high-level programming tools. An example is the Birmingham SIM_AGENT toolkit, which is unusual in avoiding commitment to any particular architecture.

LATER: raise the toolkit to a higher level, by creating a design formalism that can be compiled into something like the existing formalism.

Will need more powerful graphical extensions.

THE SIM_AGENT TOOLKIT

Built on Pop-11 (a Lisp-like language, with a Pascal-like syntax) extended with:

- **OBJECTCLASS (like CLOS - by Steve Knight, HP labs)**
- **POPRULEBASE - an unusually powerful forward chaining production system interpreter**
- **A scheduler and some default classes and methods**

(It runs in the POPLOG environment, which includes graphics, control panels, Prolog, ML, Lisp.)

CLASSES PROVIDED

Two default classes SIM_OBJECT and SIM_AGENT

Agents have an internal architecture consisting of a collection of rulesets and databases, and some default message sending and receiving methods. A 'module' in an agent is a set of rules with one or more associated databases. Linked modules share databases.

Each ruleset for each agent class has an associated 'resource' limit.

Agents have sensors and action procedures defined by methods.

A collection of default methods is provided for the two main classes. Users can define new classes and redefine the methods for those classes.

(New methods can invoke old methods, or shadow them.)

KEY IDEAS

• Behaviour is controlled by a scheduler which repeatedly "runs" all the objects and agents, in simulated time-slices.

• The scheduler uses a two-pass strategy. In each time-slice the first pass enables sensor methods to be run, and all internal processing, which may generate new external actions. In the second pass the external actions for each agent are executed.

• The world is a collection of objects, some of which may be passive (ditches, walls), some active; and active agents may be more or less autonomous and more or less intelligent. The rules for "passive" agents define the physics of the world.

• The object ontology provides a generic definition of an object class that has the basic features required for the scheduler to be able to run it using generic "methods" for running individual objects and updating their contents, etc.

• Users can define more specific sub-classes with different properties.

• More specific versions of the generic methods can be defined for sub-classes, and the scheduler will automatically invoke them.

AGENT INTERNALS

• Each agent is an objectclass instance, containing a collection of "externally visible" data held in object slots, and a set of internal mechanisms operating concurrently.

• Each internal mechanism is represented by a rule-based system with one or more rulesets, based on POPRULEBASE. These rulesets interact with internal databases. Two or more mechanisms may share part or all of a database. (But two agents will not share a database.)

• Each database forms a "working" memory for the condition-action rules, as well as forming a long term memory.

• The different rulesets and rulefamilies within an agent can be given different resource allocations, allowing relative speeds to be varied.

• The condition action rules support a mechanism for invoking neural nets and other subsymbolic systems.

• A subset of the data will be transferred between the local database and externally visible slots, or vice versa, from time to time, e.g. when perception occurs, messages come in, messages are sent out, actions occur, etc.

• The rulesets within a sub-mechanism may change over time, as may the individual rules within a ruleset.

• If more sophisticated reasoning or logical deduction procedures are required it would be possible to invoke prolog, or some sort of theorem prover. (A prolog process could be associated with each agent by using the CONSPROC facility)

• Initially we can have the same fixed collection of rulesets for each TYPE of agent. Changes in an individual agent's beliefs, goals and capabilities, including all forms of learning, will then be represented by changes in the agent's database.

• Different sets of rulesets can be used to model different architectures.

• Agents can be given different relative speeds of execution by giving them different values for their "sim_speed" slot.

It is intended that, with collaborators, we'll develop a set of libraries for different sorts of classes.

THE VIRTUAL TIME SCHEDULER

SIM_AGENT provides a scheduler which 'runs' objects in a virtual time frame composed of a succession of time slices.

It uses Objectclass methods that can be redefined for different sub-classes of agents without altering the scheduler.

The default 'run' method gives every agent a chance to do three things in each time-slice:

- **sense its environment**
- **run internal processes that interpret sensory data and incoming messages, and manipulate internal states**
- **produce actions or messages for other agents**

After doing that for each agent the scheduler uses default methods to:

- **transfer messages between agents**
- **perform the actions for each agent**

So each agent's sensory processes and internal processes run with the 'external' world in the same state in the same time-slice.

The resource limits associated with each ruleset can be varied, to allow us to explore the effects of speeding up or slowing down different internal modules relative to the speed with which things change in the environment.

This will help us evaluate the need for meta-management mechanisms.

CONCLUSION

We need collaborative investigation of many types of architecture.

This involves:
AI, Alife, Biology, Neuroscience, Psychology, Psychiatry, Anthropology, Linguistics, Philosophy, etc.

Abandoning the rationality requirement has important consequences.

People often need professional help, but the professionals don't always understand normal functioning, and therefore cannot account for deviations from normality, nor provide help reliably (except in the case of clearly defined physical and chemical abnormalities which can be remedied by drugs or surgery).

Similar possibilities arise for sufficiently sophisticated artificial agents.

Artificial agents may also need therapy and counselling, for the same reasons as humans.

Existing human therapies may fail on them too!

Acknowledgements and Notes

Work reported here has been supported at various times by the UK Joint Council Initiative, The Renaissance Trust, DRA Malvern, and the University of Birmingham. We have benefited from interactions with many research students and staff at Birmingham, in the Schools of Computer Science and Psychology.

The toolkit was developed jointly with Riccardo Poli, with contributions from Darryl Davis, Brian Logan, and several collaborators at Sussex and DRA Malvern.

It is described in
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

and in

http://www.cs.bham.ac.uk/~axs/cog_affect

Some pointers to related work can be found in

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

References

[Den78] D. C. Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, Cambridge, MA, 1978.

[McC79] 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. (Also accessible at http://www-formal.stanford.edu/jmc/ascribing/ascribing.html).

[McC95] J. McCarthy. Making robots conscious of their mental states. In *AAAI Spring Symposium on Representing Mental States and Mechanisms*, 1995. Accessible via http://www-formal.stanford.edu/jmc/consciousness.html.

[McD81] D. McDermott. Artificial intelligence meets natural stupidity. In John Haugeland, editor, *Mind Design*. MIT Press, Cambridge, MA, 1981.

[New82] A. Newell. The knowledge level. *Artificial Intelligence*, 18(1):87–127, 1982.

[SC81] A. Sloman and M. Croucher. Why robots will have emotions. In *Proc 7th Int. Joint Conf. on AI*, Vancouver, 1981.

[Slo93] A. Sloman. The mind as a control system. In C. Hookway and D. Peterson, editors, *Philosophy and the Cognitive Sciences*, pages 69–110. Cambridge University Press, 1993.

[Slo94a] A. Sloman. Explorations in design space. In *Proceedings 11th European Conference on AI*, Amsterdam, 1994.

[Slo94b] A. Sloman. Semantics in an intelligent control system. *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering*, 349(1689):43–58, 1994.

[Slo95a] A. Sloman. Exploring design space and niche space. In *Proceedings 5th Scandinavian Conference on AI, Trondheim*, Amsterdam, 1995. IOS Press.

[Slo95b] A. Sloman. Musings on the roles of logical and non-logical representations in intelligence. In Janice Glasgow, Hari Narayanan, and Chandrasekaran, editors, *Diagrammatic Reasoning: Computational and Cognitive Perspectives*, pages 7–33. MIT Press, 1995.

[Slo96a] Aaron Sloman. Actual possibilities. In Luigia Carlucci Aiello and Stuart C. Shapiro, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*, pages 627–638. Morgan Kaufmann Publishers, 1996.

[Slo96b] Aaron Sloman. Towards a general theory of representations. In D.M.Peterson, editor, *Forms of representation: an interdisciplinary theme for cognitive science*. Intellect Books, Exeter, U.K., 1996. ISBN: 1-871516-34-X.

[WSB96] 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. Available at URL ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect in the file Wright_Sloman_Beaudoin_grief.ps.Z.