Synthetic Minds

Aaron Sloman & Brian Logan
School of Computer Science
The University of Birmingham
Birmingham, B15 2TT, UK
Phone: +44 121 414 4775
Fax: +44 121 414 4281
A.Sloman@cs.bham.ac.uk
B.S.Logan@cs.bham.ac.uk
www.cs.bham.ac.uk/ ~axs/cog_affect

Abstract
This paper discusses conditions under which some of the “higher level” mental concepts applicable to human beings might also be applicable to artificial agents. The key idea is that mental concepts (e.g. “believes”, “desires”, “intends”, “mood”, “emotion”, etc.) are grounded in assumptions about information processing architectures, and not merely Newell’s knowledge-level concepts, nor concepts based solely on Dennett’s “intentional stance.”

1 Describing synthetic agents

McCarthy [McC79, McC95] 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 paper 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 interrelated concepts namely the concepts that describe the states and processes able to occur in the architecture.

An 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. That possibility 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

The relationship between mental concepts and the underlying architecture can be compared with the way in which a new theory of the architecture of matter generated the table of possible elements: the periodic table.

Within the framework of the atomic theory of matter developed during the last two centuries, 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.

Similarly a good theory of the architecture of a type of agent is likely to show the need for revisions and extensions of our existing theory of types of states in such agents. Compare approaches that start by defining types of states and then try to derive architectures.

Mentalistic concepts applicable to artificial agents

It is often convenient to describe a machine as "choosing", "exploring", "deciding", "inferring", etc. The states and processes referred to are *intentional*, since they have semantic contents.

In some cases it may be useful also to describe such systems as "believing", "wanting", "preferring", "enjoying", "disliking", "frightened", "angry", "relieved", "delighted".

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 if applying mentalistic terms to them is not to be misleading, like the over-enthusiastic use of words like "goal" and "plan" in some AI publications, criticised by McDermott [McD81].

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. This anti-behaviourist assumption has a long history and will not be defended here.

2 Why use mentalistic language?

We shall need mentalistic descriptions for synthetic agents (a) because of marketing requirements, (b) because such descriptions will be irresistible and (c) because no other vocabulary will be as useful for describing, explaining, predicting capabilities and behaviour. ((c) provides part of the explanation for (b).)

E.g. descriptions in terms of physical processes, or the programming language level data-structures and algorithms will not be useful for those who have to interact with the agents, however useful they are for developers and maintainers. This is analogous to the fact that interacting with people is difficult if the only way you can think about them is in terms of their internal physiological states.

So, instead of trying to avoid the use of mentalistic language, which will be self-defeating, 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'.

Unlike Dennett and Newell ...

This differs from the approach of Dennett [Den78] who recommends the "intentional stance" in describing sophisticated robots, as well as human beings. This stance presupposes that the agents being described are rational: otherwise their behaviour provides no basis for inferring beliefs, desires, intentions, etc.

Our stance also differs from the approach of Newell [New82] who recommends the use of "knowledge level", which also presupposes rationality.

... We use an "information level" design stance

Our claim is that mentality is concerned with an "information level" architecture, close to the requirements specified by software engineers. This is a version of the 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 "holistic" intentional stance permits only talk about what **the whole** agent believes, desires, intends, etc. Information level design descriptions also allow us to talk about various semantically rich **internal** information stores, motive databases, state transitions that are possible for internal information items (e.g. being generated, evaluated, adopted, rejected, stored for future consideration, interrupted, suspended, reactivated, modified, destroyed, matched against other items, etc.)

3 Rationality is not an absolute requirement for mentality

The mechanisms in such an architecture need be neither rational nor irrational: even though they acquire information, evaluate it, use it, store it, etc. [Slo94b]. Some of the processes are neither rational nor irrational because they are **automatic**.

We claim that does not prevent them being concerned with semantic information (including internal references: such as one internal structure that is used by the machine to describe the relationship between two other structures, for instance a history of changes in plans which may be useful in preventing looping and other wasted actions during planning).

There is no commitment at this stage regarding the **form** used to encode or express information. It may include logical databases, procedures, image structures, neural nets or in limiting cases physical representations, such as curvature of a bimetallic strip representing temperature. (For more on this see [Slo95b, Slo96a, Slo96b].)

At this level we can begin to explain what mental states are in terms of 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. Thus semantic content is already present, without full-blown intentionality or rationality.

By describing a variety of functions using the "design stance" at the information level, and showing how they implement mental states and processes, we provide a richer and deeper explanatory framework than the intentional stance.

4 Emergent states and processes

Not all states require specific mechanisms in the architecture. A computing system that is "overloaded" does not have an "overloading" mechanism. Rather that's a feature of the interaction of many different mechanisms all of which have functions other than producing overload. Similarly with many mental states, e.g. emotions.

If the system also has the ability to monitor its own states and processes a new variety of descriptions becomes applicable, including new forms of self control, learning of concepts for self-description, etc.

In particular, the phenomena often described by philosophers and others as involving "qualia" may be explained in terms of high level control mechanisms with the ability to attend to many internal states and processes including internal intermediate structures produced during the processing of sensory information.

The objects of such self-monitoring processes may be virtual machine states rather than internal physical or physiological states. Software agents able to inform us (or other artificial agents) about their own internal states and processes may need similar architectural underpinnings for qualia.

This need be no different from the mechanisms underpinning 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).

Another example follows.

5 Example: What is required for carelessness?

Describing X as "working carelessly" implies

(a) that X had certain capabilities relevant to the task in hand,

(b) that X had the ability to check and detect the need to deploy those capabilities,

(c) that the actual task required them to be deployed (e.g. some danger threshold was exceeded, which could have been detected, whereupon remedial action would have been taken),

(d) that 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).

This illustrates how familiar mentalistic descriptions can presuppose a design architecture.

The presuppositions for "working carefully" are similar to those for working carelessly. Something that is incapable of being careless cannot be careful.

6 Talking about artificial agents

Our claim is that when people use mentalistic language to describe themselves or other humans they implicitly presuppose that there are various coexisting interacting subsystems with different functional roles, for instance, perceptual subsystems, various types of memory, various skill stores, motivational mechanisms, various problem solving capabilities.

There is no reason why we should not transfer these predicates to artificial agents, if they have appropriate architectures.

7 How to make progress

A task for agent theorists is to devise a more accurate and explicit theory of the types of architecture to be found in human minds (and others) 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. Layered architectures may be important.

We conjecture that human-like agents need an architecture with at least three layers (see figures below):

- 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 in some other animals (e.g. cats, monkeys).
- 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?)

8 Architectural layers and types of emotions

These layers account for different sorts of mental states and processes, only some of which are shared with other animals [WSB96].

Many disagreements about the nature of emotions seem to be based on a failure to grasp that there are different concepts of emotionality which presuppose different architectural features, not all of which are shared by some of the animals studied by emotion theorists.

In particular, it is not always noticed that there are different sorts of **emotional** states and processes based on the different layers, e.g.:

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

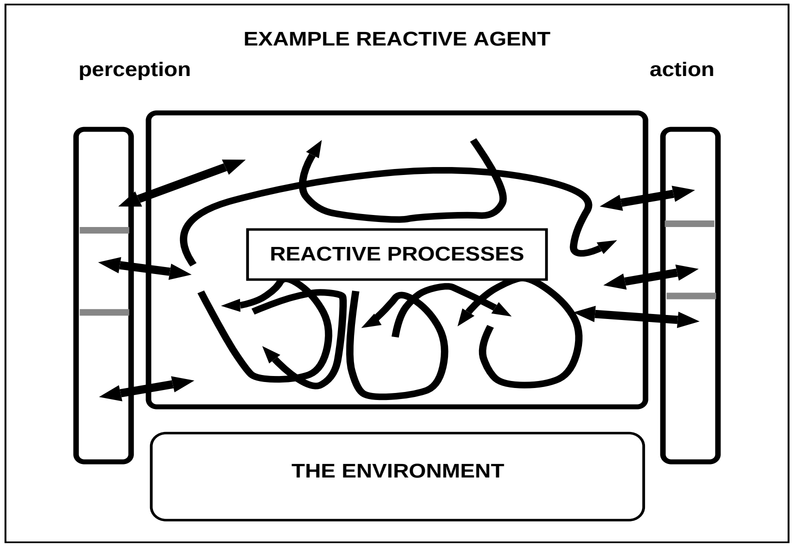
(2) emotional states (like being anxious, apprehensive, relieved, pleasantly surprised) which depend on the existence of the deliberative layer, in which plans can be created and executed,

(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.

The second class of states depends on abilities that appear to be 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.

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.

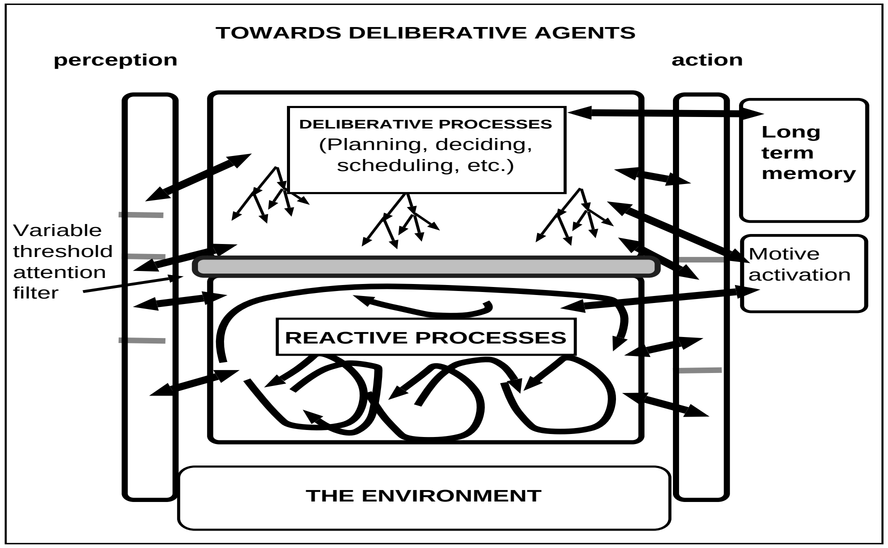
9 Reactive agents



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?).

10 Combining reactive and deliberative layers



In a deliberative mechanism

- New plans may be constructed
- Options are explicitly evaluated before selection
- Re-usable mechanisms and space are dynamically allocated, making many processes inherently serial
- Learnt skills can be transferred to the reactive layer (if there's spare capacity)
- 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.

11 The need for self-monitoring (meta-management)

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

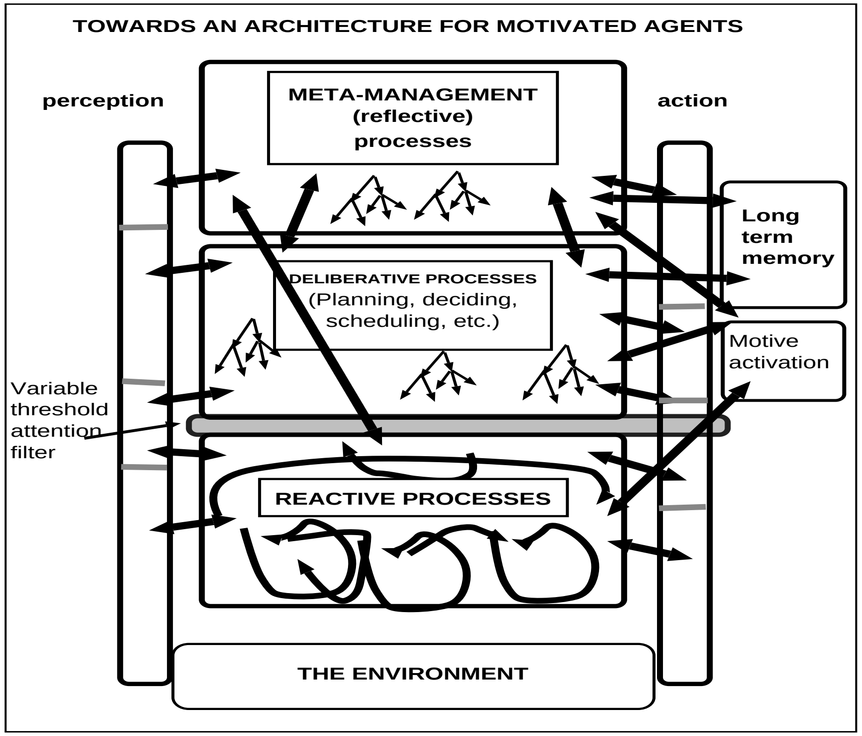
However, deliberative mechanisms with evolutionarily determined strategies for planning, problem solving, decisions making, evaluating, can be too rigid.

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,
- Notice 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 ("A little to the left of where the hillside intersects the tree trunk"), or using drawings and paintings to communicate about how things look.

Meta-meta-management may not be needed if meta-management mechanisms are recursive (i.e. partly self-applicable)!

12 Towards multi-layered autonomous (reflective) agents



Generic functions of internal self-monitoring, "meta-management" processes could include:

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

13 There is no unique architecture

Different kinds of meta-management are likely to be found in different animals.

Many architectures are needed for different sorts of organisms or artificial agents.

Even humans differ from one another. Architectures may differ between between human children, adolescents, adults and senile adults. Perhaps there are also culturally determined differences in architectures.

Meta-management and deliberative mechanisms permit cultural influences via the absorption (and transmission) of new concepts structured descriptions, and rules, norms and evaluation criteria).

Similarly, 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.

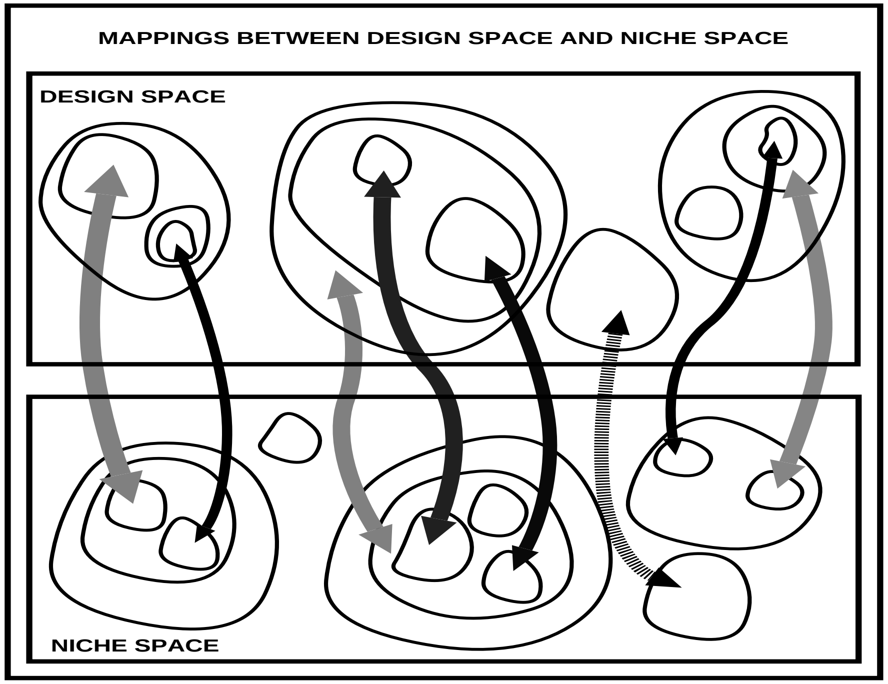
If these conjectures are correct, then designers of synthetic agents need to be aware of the evolutionary pressures that led to these layers in human beings. Perhaps they are also required for certain classes of sophisticated artificial agents, whether robots or software agents.

In that case, there may be some unanticipated consequences of these design features [SC81].

Analysing these possibilities is hard. By developing a theory of a space of possible architectures [Slo93, Slo94a, Slo94b, Slo95a] we provide a framework for more precise specifications of alternative families of mentalistic concepts.

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

14 DESIGN SPACE and NICHE SPACE



Notes

- A niche is a set of requirements
- A design is a set of specifications
- Mappings are not unique: there are always trade-offs
- 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.

15 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 (as explained in [Slo95a].

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.

However, it may turn out that for many architectures, including some found in nature and in artificial agents, normal modes of description may not be appropriate. For those we'll need to develop new systems of concepts and explanatory principles. (Can a goldfish long for its mother, and if not why not?)

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.

16 Conclusion

We have attempted to sketch a framework within which collaborative investigation of many types of architecture of varying degrees of sophistication, with varying mixtures of information-processing capability may be possible, including AI, Alife, Biology, Neuroscience, Psychology, Psychiatry, Anthropology, Linguistics and Philosophy.

This depends on identifying an important level of analysis to which the design stance can be applied: the information processing level. This is close to but different from Dennett's intentional stance and Newell's knowledge level, partly because it is concerned with mechanisms for which considerations of rationality do not arise.

Moreover, any general theory of agents should not focus on rationality as a central criterion of agency. It could rule out humans!

Even folk psychology makes allowance for impulses, obsessions, addictions, memory lapses, various kinds of carelessness, temporary misjudgements of relative importance, and so on. Professional counsellors and therapists have additional specialised ways of categorising mental states and processes without presupposing rationality (though which of them will survive creation of good theories about the underlying architecture is an open question).

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 [SC81]. Existing human therapies may fail for the same reasons.

We need all these different types of exploration to proceed in parallel, including philosophical analysis, psychological and neurophysiological studies of humans and other animals, experiments with a variety of working models of agents, and evolutionary processes that might throw up types of architectures that we would not otherwise think of.

This may force us to invent new concepts for describing some sorts of synthetic minds.

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.

A toolkit written in Pop-11, developed jointly with Riccardo Poli for exploring a variety of types of agent architectures and doing evolutionary experiments 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.