Presented at I3 Spring Days Workshop on
Behavior planning for life-like characters and avatars
Sitges, Spain, March 1999

Beyond Shallow Models of Emotion

Aaron Sloman
School of Computer Science
The University of Birmingham
Birmingham, B15 2TT, UK
http://www.cs.bham.ac.uk/˜axs/

Abstract:
There is much shallow thinking about emotions, and a huge diversity of definitions of "emotion" arises out of this shallowness. Too often the definitions and theories are inspired either by a mixture of introspection and selective common sense, or by a misdirected neo-behaviourist methodology, attempting to define emotions and other mental states in terms of observables. One way to avoid such shallowness, and perhaps eventually achieve convergence, is to base concepts and theories on an information processing architecture, which is subject to various constraints, including evolvability, implementability, coping with resource-limited physical mechanisms, and human-like functionality. Within such an architecture-based theory we can distinguish (at least) primary emotions, secondary emotions, and tertiary emotions, and produce a coherent theory which explains a wide range of phenomena and also partly explains the diversity of theories: most theorists focus on only a subset of types of emotions.

Keywords: affect, architecture, artificial intelligence, cognitive science, deliberative, emotion, evolution, intelligence, meta-management, mind, reactive

1. Introduction

The study of emotion in cognitive science and AI has suddenly become very fashionable, with a rapidly growing number of workshops, conferences and publications on the topic. Of course, it is not a new topic, even in AI, as shown by Simon's important contribution over 30 years ago [14], and various papers nearly 20 years ago in IJCAI'81 including my first paper on this topic [15]. Although there are some excellent surveys of issues concerning emotions (e.g. [7, 11–13]), it is difficult for newcomers to the field to achieve a balanced overview, and in consequence there is a growing tendency to present simplistic AI programs and robots as if they justified epithets like "emotional", "sad", "surprised", etc. This is similar to the tendency, lambasted long ago by McDermott in [8], to use terms like "goal", "plan", "learn", simply because there are procedures or variables with these names in a program.

A typical manifestation of such shallowness is having one or more emotional state variables either with boolean values that can be toggled or with a numerical or "qualitative" range of values for each variable. Such models are hopelessly inadequate in accounting for typical human social emotions which are rich in semantic content, for instance being infatuated, or feeling humiliated because some silly mistake you made was pointed out by a famous person in a large public lecture.

2. Shallow models are not all bad

Shallow models may not matter if they have a limited purpose which is made clear, e.g. to entertain, or to teach programming, or to model some limited aspect of control of posture or facial expression, etc. I have a very shallow model in which simulated mobile robots can be in states described as glum, surprised, neutral or happy, but this is nothing more than an elementary teaching tool. Students play with and extend it in order to learn agent programming techniques. In the near future, there will probably be a growing use of very shallow models of emotion in computer entertainments. There is nothing wrong with that, if they are successful at entertaining. However that does not necessarily make them plausible models of human or animal emotions. They may not even be useful steps in the direction of such models.

Shallow models can sometimes play a role in the search for deeper models. Building inadequate models, and exploring their capabilities and limitations is often an essential part of the process of learning how to design more complex and more satisfactory models, as explained in [1, 19].

3. How to achieve greater depth

A desirable but rarely achieved type of depth in an explanatory theory is having a model which accounts for a wide range of phenomena. One of the reasons for shallowness in psychological theories is consideration of too small a variety of cases.

If instead of thinking only about normal adult humans we consider also infants, people with brain damage or disease, and also other animals including insects, bacteria, birds, bonobos, etc., we find evidence for myriad information processing architectures each supporting and explaining a specific combination of mental capabilities. Yet more possible architectures, each supporting a collection of possible states and processes can be found in robots, software systems and machines of the future!

Thus concepts describing mental states and processes in one animal or machine may be inappropriate when describing another. Likewise, concepts relevant to normal adult humans may be inappropriate for new-born infants, victims of Alzheimer's disease, or an entertaining robot which can be made to *look* happy, annoyed, surprised, etc.

Although human adults seem to be innately programmed to attribute all sorts of mental states to infants, in fact infants may be incapable of having some of them. For instance, a newborn infant may be incapable of feeling humiliated if it lacks the required architecture. It may even be incapable of feeling pain in the same way as an adult, despite displaying compelling external symptoms.

It often goes unnoticed that much of what poets and novelists say about us, and what we say about our friends and ourselves when gossipping or discussing our interests, loves, hopes, fears and ambitions, implicitly presupposes that humans are essentially information processing systems. E.g. when poets distinguish *fickle liking* which is easily diminished by new information and *deep love* which is not, they implicitly presuppose that new information can have effects on powerful information-based control states.

By considering possible descriptive and explanatory concepts generated by a *virtual machine information processing architecture* we obtain a broader and deeper explanatory theory than is normally found in philosophy, psychology or social science. Of course, such a theory should satisfy empirical constraints including evolvability, implementability in neural mechanisms, resource limits, etc.

A comprehensive theory of emotions and other mental states requires a survey of types of information processing architectures covering humans of various types, other animals, future robots and software agents. For each type of architecture we can precisely define the sorts of states and processes it supports, and then we can formulate and, perhaps begin to answer, far more precise questions about which agents are capable of having which sorts of emotions, experiences, thoughts, and so on.

A proper understanding requires comparative analysis of possibilities and trajectories in design space and niche space, as outlined in [20, 23]. We understand a particular architecture better if we know what differences would arise out of various sorts of design changes: which capabilities would be lost and which would be added. We also have a deeper understanding of the architecture if we can see what sorts of pressures and trade-offs led

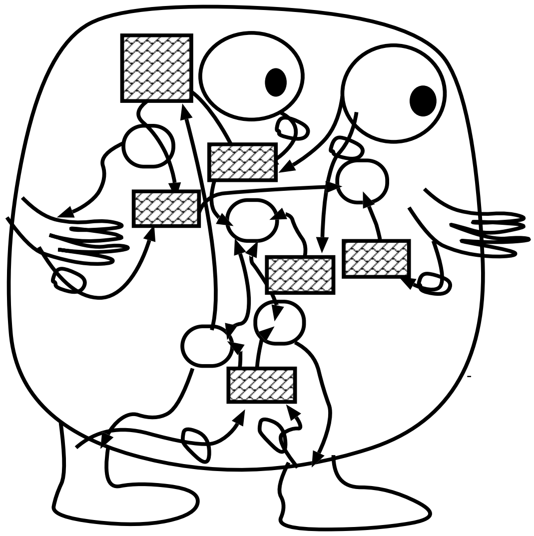

Fig. 1. *An unstructured mess?*
Any observed behaviour might be produced by an unintelligibly tangled and non-modular architecture. (*Rectangles represent information stores and buffers, ovals represent processing units, and arrows represent flow of information.*)

to its evolution, and how it might develop or evolve in future.

This involves going beyond the majority of AI projects or psychological investigations insofar as it requires us both to consider designs for *complete* agents and also to do *comparative* analysis of different sorts of designs.

4. Constraints on theorising

Discovering the architecture of a complex system we have not designed ourselves is very difficult. No amount of observation of the behaviour of any animal or machine can determine the underlying architecture, since in principle any lifelong set of behaviours can be produced by infinitely many different information processing architectures, including totally unstructured, unintelligible, "flat", multi-component architectures, as suggested in Figure 1.

Decompiling information gleaned from invasive or non-invasive observation of internal physical structures is just as hard, e.g. if we don't even know at what physical level most of the architecture is implemented. Do neurons or molecules do most of the information processing?

We can best constrain our theories by combining a number of considerations which I have discussed a greater length in [23, 26], such as: (1) trade-offs that can influence evolutionary developments, (2) what is known about our evolutionary history, (3) what is known about

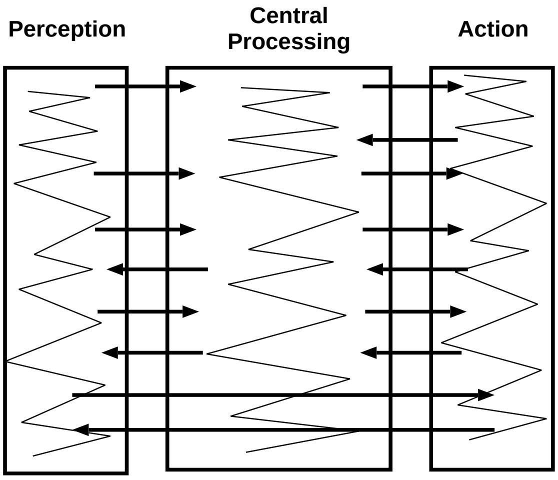

Fig. 2. *A triple tower model* (*based on Nilsson*)

Intelligent organisms and robots require perceptual mechanisms and action mechanisms of varying degrees of sophistication. In general there are also more central processing mechanisms. The boundaries between the "towers" need not be very sharp, especially where there is rich two-way information and control flow.

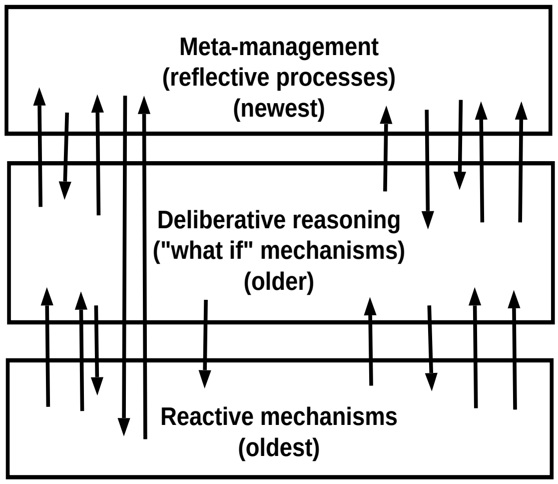

Fig. 3. *The triple layer model*

There is good reason to believe that early organisms, like some existing organisms, were totally reactive, and that deliberative and meta-management layers evolved later. Adult humans appear to have all three types of processing, which is probably rare among animals. The three layers operate concurrently, and do not form a simple dominance hierarchy. Imagine this model superimposed on Figure 2.

human and animal brains and the effects of brain damage, (4) what we have learnt in AI about the scope and limitations of various information processing architectures, mechanisms and representations, (5) introspective evidence, such as my knowledge that I considered and evaluated alternative ways of travelling to the I3 Spring Days conference before buying tickets.

But our theories will still remain *conjectures* for a long time to come. At least we can show that some conjectures are better than others, if we take a broad enough view of what needs to be explained.

5. Towards a sketch of a theory

Nilsson [10] has listed some reasons for supposing that intelligent systems can be analysed in terms of the "triple tower" model depicted in Figure 2, which approximately separates perceptual mechanisms, central processing mechanisms and action mechanisms. He calls the central tower the "model tower", though this label may be too restrictive for the range of functions sketched below. The triple tower model is mainly a result of functional analysis combined with observation of existing organisms.

Another breakdown of information processing functionality comes from both functional and evolutionary considerations. This is the triple layer model sketched in Figure 3, and discussed at greater length in previous papers (e.g. [21, 24, 22, 26, 16]). These three levels are different from the three discussed by Nilsson in chapter 25 of [10], though there is some overlap.

If the three layers and the three towers are superimposed we arrive at an architecture where perceptual mechanisms have several layers with different kinds of sophistication, which evolved at different times to fit in with the requirements of the different central layers. Likewise the action mechanisms may have different level of sophistication supporting different sorts of functionality which evolved at different times.

All of this is part of a conjectural theory of a normal adult human information processing architecture based on evidence of many kinds from several disciplines, and the sorts of constraints on evolvability, implementability and functionality mentioned above.

According to this theory:

(a) Evolution, like engineers, found that (partly) modular designs are essential for defeating combinatorics in the search for solutions to complex problems (with only 4,000,000,000 years and one biosphere on an earth-sized planet available).

(b) Human information processing makes use of (at least) three different concurrently active architectural layers, a reactive layer, a deliberative layer, and a meta-management layer which evolved at different times, which we share with other animals to varying degrees, along with various additional supporting modules such as motive generators, "global alarm" mechanisms and long term associative storage mechanisms. The different layers and supporting mechanisms may have evolved from purely reactive mechanisms by means of the typical evolutionary trick of making another copy of an existing mechanism and then gradually transforming the functions of the new copy. This almost certainly happened several times in the evolution of brains.

(c) Reactive systems may be very complex, and powerful, especially if internal reactions can be chained together and can cause modification of internal states which trigger or modulate other reactions. I do not claim that deliberative or meta-management mechanisms provide behavioural capabilities that could not *in principle* be provided by purely reactive mechanisms. Rather I have argued elsewhere that achieving the same functionality by purely reactive means would have required a far longer period of evolution with more varied circumstances, and a far larger brain to store all the previously evolved reactive behaviours. The time and brain size required for a purely reactive human-like system are probably too large to fit into the physical universe. Some people who argue in favour of purely reactive systems do not consider the trade-offs involved in these resource issues. Merely showing that in principle reactive systems suffice proves nothing about what can work in practice.

(d) Reactive, deliberative and reflective layers support different classes of emotions found in humans and other animals, including the primary and secondary emotions discussed by Damasio and Picard [4, 13], and the tertiary emotions I have discussed in criticising their work [22, 25].
(i) the reactive layer, including a global alarm mechanism, accounts for *primary* emotions (e.g. being startled, frozen with terror, sexually aroused);
(ii) the deliberative layer supports *secondary* emotions like apprehension and relief which require "what if" reasoning abilities (these are semantically rich emotions);
(iii) a meta-management (reflective) layer supports not only control of thought and attention but also loss of such control, as found in typically human *tertiary* emotions such as infatuation, humiliation, thrilled anticipation of a future event. (This layer is also crucial to absorption of a culture and various kinds of mathematical, philosophical and scientific thinking.)

All the layers are subject to interference from the others and from one or more fast but stupid partly trainable "global alarm" mechanisms (e.g. spinal reflexes of various sorts, the brain stem, the limbic system including the amygdala, etc.)

(e) A more fine-grained analysis of types of processes that we tend to call "emotions" in humans would show that the above three-fold classification into primary, secondary and tertiary emotions is somewhat superficial. For instance, there are different ways emotions can develop over time, and the three-fold distinction does not say anything about that. A short flash of anger or embarrassment which quickly passes is very different from long term brooding or obsessive jealousy or humiliation which gradually colours more and more of an individual's mental life.

(f) Perceptual and motor systems are also layered: the different layers evolved at different times, act concurrently, and have different relationships to the "central" layers. E.g. deliberative mechanisms make use of high level characterisations of perceived states, e.g. seeing a bridge as "rickety" or an ornament as "fragile". Using some of Gibson's ideas, this can be described as perception of abstract affordances.

(g) Analysing ways in which components of such an architecture might bootstrap themselves, develop, reorganise themselves, acquire and store information, or go wrong, will provide far richer theories of learning and development than ever before.

(h) The three layers account for different cognitive and affective states, as well as different possible effects of brain damage, and other abnormalities. For instance, some aspects of autism seem to involve malfunctioning or non-functioning higher level perceptual mechanisms (as suggested in [17]).

(i) A multi-layered architecture of the sort proposed could give robots various kinds of human-like mental states and processes, including *qualia* arising out of inward focused attention. As science fiction writers have noted, this might lead some robots to re-discover philosophical confusions about consciousness. Software agents could have similar capabilities. However, detailed differences in physical embodiments and virtual machine architectures could entail many kinds of minor differences in the mental states of which they are capable. This is no different in principle from the fact that mental states possible for adults and children are different, or for males and females, or humans and cats.

Many doubt these claims about robots because they see the limitations of existing computer-based machines and software systems and cannot imagine any ways of overcoming these limitations. They do not realise that we are still in the early stages of learning how to design information processing systems. (Claiming that computers will be ever more powerful is not enough to allay these doubts: we also need deep analysis of the concepts used to express the doubts.)

6. Alternatives in design space

Although the above theory includes a sketch of an architecture for human-like intelligent systems, there is no suggestion that this is the only sort of intelligence. 'Intelligence', like 'emotion', is a *cluster concept*, referring to a variable cluster of capabilities, and admitting a wide variety of types of instances, with no sharp

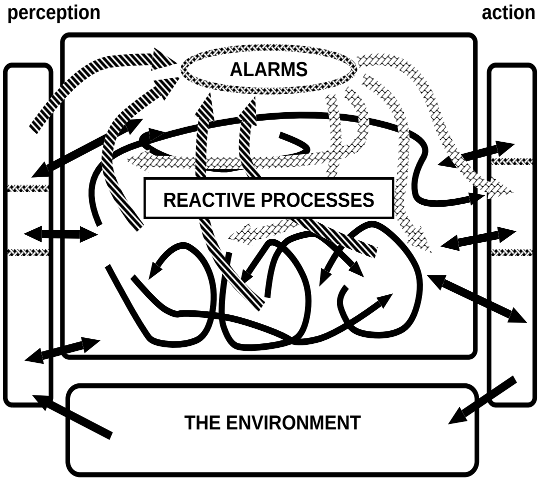

Fig. 4. *A reactive system with global alarms.*
When reactive systems are so complex and sophisticated that they can introduce significant delays between sensing and acting, it may be useful to have a more 'stupid' pattern-directed alarm system, with inputs from everywhere and outputs going to all parts of the system, which can take control when emergencies or urgent opportunities are detected.

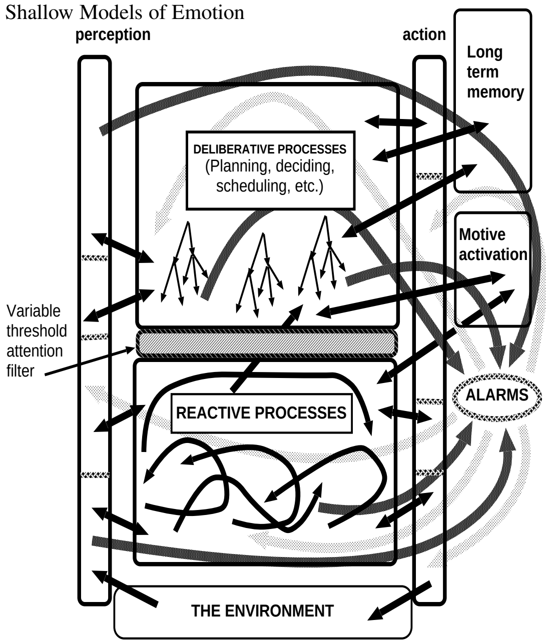

Fig. 5. *A hybrid architecture with global alarms.*
Reactive and deliberative mechanisms may sometimes be dominated by control signals from a global alarm system.

boundaries. In particular, animals (and perhaps humans) exist with different subsets of the full array of mechanisms described above, and within those mechanisms considerable variation is possible.

For example, many insects appear to be capable of remarkable achievements based entirely in complex collections of purely reactive mechanisms, such as termites constructing their "cathedrals", with air conditioning, nursery chambers and other extraordinary features.

So I am not denying that there can be organisms (and robots) which are purely reactive, or which combine a reactive mechanism with a separate global alarm system, as in Figure 4.

More sophisticated organisms have both a reactive and a deliberative layer, providing "what if" reasoning capabilities, as illustrated in Figure 5. Such mechanisms provide the ability to construct specifications of hypothetical past or future situations and to reason about them. Many writers, including Craik [3] as long ago as 1943, have pointed out that such abilities may increase biological fitness.

It seems that some other animals besides humans have deliberative mechanisms though they vary enormously in their richness and flexibility. For instance, how effective such capabilities are, will depend on a number of factors including the type and size of re-usable short term working memory, the type of representational mechanisms available, the type and size of the trainable associative memory which can store generalisations about the environment, and so on.

The deliberative layer might have evolved as a result of a mutation which at first led to the copying of a trainable associative memory in a purely reactive system. After that, the new copy might have gradually evolved, along with other mechanisms, to provide the ability to answer questions about "what would happen if" instead of "how shall I react now". Making good use of such a "what if" reasoning capability requires being able to store generalisations about the environment at an appropriate level of abstraction to allow extrapolation beyond observed cases. This in turn could generate evolutionary pressure towards perceptual systems which include higher level abstraction mechanisms. All this is, of course, highly speculative, and needs to be tested empirically, though it is consistent both with what is known about evolutionary mechanisms and with the at least partly modular structure of the brain.

More generally, within this framework we can see a need for a generalisation of Gibson's theory of perceptual affordances [6] (contrasted with Marr's theory of vision in [17]) to accommodate different perceptual affordances for different components in the more central processing mechanisms. This requires the sharing of sensory resources between concurrently active subsystems, and can generate conflicts, as discussed in [18].

Deliberative capabilities bring their own problems, such as how they should be controlled, how different deliberative strategies should be selected or interrupted, how they should be evaluated and modified. For this

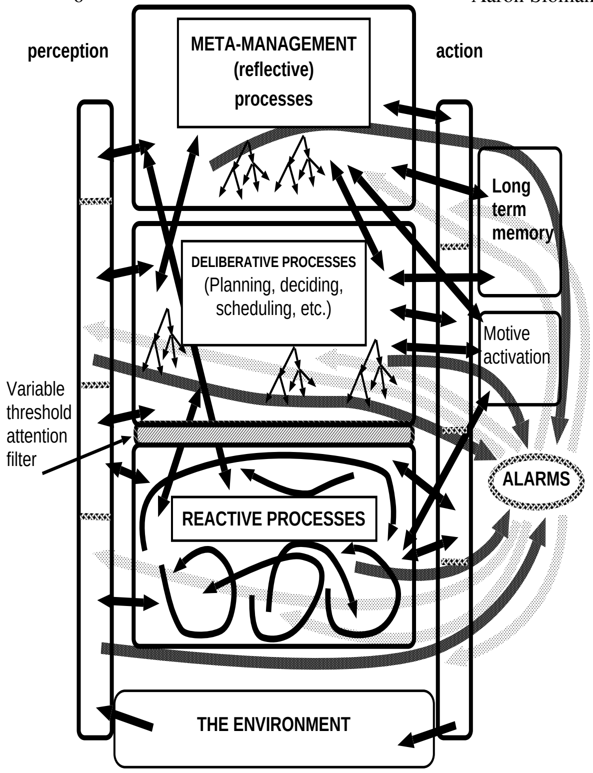

Fig. 6. *Adding a meta-management layer.*
The meta-management layer provides the ability to attend to, monitor, evaluate, and sometimes change internal processes and strategies used for internal processes. However, all the layers and the alarm system(s) operate concurrently, and none is in total control.

purpose and others, it seems that an even smaller subset of animals, including humans, have evolved a third architectural layer providing the ability to direct attention *inwardly* and to monitor, evaluate, and in some cases modify what is happening internally. Luc Beaudoin first drew my attention to some aspects of the need for this layer, and called it meta-management. Some of the requirements were analysed in his PhD thesis [2].

Earlier papers (e.g. [27]) have discussed some of the ways in which this theory accounts for distinctively human emotions such as grief, infatuation, excited anticipation, humiliation, involving partial loss of control of attention. We used to call these emotions "perturbances", but now refer to them as tertiary emotions, to distinguish them from the primary and secondary emotions discussed by Damasio and others.

Since these tertiary emotions (perturbances) involve loss of control of attention, and you cannot lose what you have not got, only an organism which has something like meta-management capabilities can get into such states. This does not mean that all humans have this capability. New born infants, people with degenerative brain disease or brain damage, may lack such capabilities.

7. Are emotions required for intelligence?

It is clear that local reflexes and global alarm mechanisms can be useful in organisms or machines which sometimes require very rapid reactions to occur faster than normal processes of perception, reasoning, deliberation, and planning. Such reactions can produce simple and obvious effects such as freezing, fleeing, producing aggressive sounds or postures, pouncing on prey, sexual responses, and more subtle internal effects such as attention switching and "arousal" which might involve different kinds of information processing. Because these reactions often need to happen very quickly they can be triggered by a relatively stupid, but trainable, pattern recognition system.

Many human emotions seem to involve the operation of such mechanisms. These and other emotions are connected with resource-limits in more "intelligent" subsystems. If those systems could operate faster, and with more complete information, it would not be necessary for more "stupid" mechanisms to override them.

Damasio (in [4]) pointed out that certain kinds of frontal lobe damage can simultaneously remove the ability to have certain classes of emotions and also undermine the ability to achieve high level control of thought processes required for successful management of one's life. Pending further investigation of details, this gives some support for the claim that there are classes of emotions, referred to as "tertiary emotions" above, which depend on mechanisms that are concerned with high level management of mental processes.

Damasio argued from this that emotions are a *requirement* for intelligence, and since then the argument has been repeated many times: it has become a sort of *meme*. However, the reasoning is fallacious, as I have argued in [22, 25]. The brain damage in question might merely have disabled some mechanisms involving control of attention, required *both* for tertiary emotions and for management of thought processes. It doesn't follow that emotions somehow contribute to intelligence: rather they are a side-effect of mechanisms that are required for other reasons, e.g. in order to overcome resource limits as explained above.

Here's an example of similarly fallacious reasoning that nobody would find convincing. Operating systems which support multiple concurrent processes are extremely useful, but they can sometimes get into a state where they are "thrashing", i.e. spending more time swapping and paging than doing useful work. If some damage occurred which prevented more than one process running at a time that would prevent the thrashing, and remove the useful benefits of multi-processing. It doesn't follow that a thrashing mechanisms is required to produce useful operating systems. In fact, by adding more memory and CPU power, thrashing can be reduced and performance enhanced. Likewise, it is possible for mature humans to learn strategies for avoiding emotions, and this can often improve the quality of their lives and the lives of people they live with or work with.

I am not arguing that all emotions are undesirable or dysfunctional, merely refuting a fallacious argument. There are many emotions that have an important biological role (e.g. sexual passion, and aggression in defending a nest), and some emotions that humans value highly, including aesthetic emotions and the joy of discovery. I also accept, as most AI researchers have accepted over many years, that there are many purely intellectual problems which require exploration of search spaces that are too large for complete, systematic, analysis. The use of heuristic pattern-recognition mechanisms is often useful in such cases, to select avenues to explore and to redirect processing. But they can operate without generating any emotions.

8. Conclusion

This paper is a snapshot of an ongoing long term multi-disciplinary research project attempting to understand the nature of the human mind and how we fit into a larger space of possible designs for biological organisms and artificial agents of many kinds.

The ideas have many links with previous work by others. Besides the connection with Simon's, Gibson's and Nilsson's ideas cited above, there are obvious links with Dennett and Minsky (e.g. [5, 9]). However there is no room for a survey of similarities and differences.

There has also not been space to explore all the implications, but one thing is very clear: we are a long way from implementing artificial systems with the full richness and complexity of the systems described here.

There are many gaps in what current AI systems can do, insofar as they are thought of as steps towards modelling human intelligence, and beyond. Existing AI systems do not yet have whatever it takes to enjoy or dislike doing something. They do not really *want* to do something or *care* about whether it succeeds or fails, even though they may be programmed to give the superficial appearance of wanting and caring, or feeling happy or sad. animal-like wanting, caring, enjoying, suffering, etc. seem to require types of architectures which have not yet been analysed.

Simulated desires and emotions represented by values for global variables (e.g. degree of "fear") or simple entries in databases linked to condition-action rules may give the appearance of emotion, but fail to address the way semantically rich emotions emerge from interactions within a complex architecture, and fail to distinguish different sorts of emotions arising out of different types of processing mechanisms within an integrated architecture.

Current AI models of other animal abilities are also limited: for example, visual and motor capabilities of current artificial systems are nowhere near those of a squirrel, monkey or nest-building bird. To understand animal comprehension of space and motion we may need to understand the differences between precocial species born or hatched with considerable independence (chickens, deer) and altricial species which start utterly helpless (eagles, cats, apes). Perhaps the bootstrapping of visuo-motor control architectures in the latter yields a far deeper grasp of space and motion than evolution could have pre-programmed via DNA. The precocial species may have much simpler visual capabilities, largely genetically determined.

There are many issues that are still unclear, and a vast number of remaining research topics. In particular it is not clear how much of this is relevant to the design of software agents inhabiting virtual machine environments only, and lacking physical bodies. Many of the human reactive mechanisms and some of their motivators and emotional responses are closely linked to bodily mechanisms and functions. E.g. if you don't have a body you will never accidentally step on an unstable rock, and you will not need an "alarm" mechanism that detects that you are about to lose your balance and triggers corrective action, including causing a surge of adrenalin to be pumped around your body.

Nevertheless events can move fast in a virtual machine world (as many system administrators fighting malicious intruders will confirm) and even pure software agents may need reactive mechanisms. Still, it is likely that the combinations required for software agents may include some architectures never found in agents with physical bodies. Whether the reverse is the case depends on whether all sorts of physical bodies and physical environments can, in principle, be simulated on sufficiently powerful physically implemented computers: an open question.

Artificial agents which do not share our deep grasp of spatial structure and motion will be limited in their ability to communicate with us. However, it is not obvious that in order to share this knowledge such agents *must* have similar bodies and processing architectures. For instance, people who have never wanted to kill someone, may nevertheless understand some of the thought processes of a murderer (a fact on which the success of many novels and plays depends). Similarly someone who has been blind from birth can understand a great deal about visual capabilities of sighted people, for instance, that colours are extended properties of 2-D surfaces, somewhat like tactile textures.

So it remains possible that some software agents which are very unlike us will be able to engage in rich communication with us, though the detailed requirements for this are still not clear.

And of course, in the meantime, teachers and designers of computer games can build many entertaining or didactic, shallow simulations which lack most of the features discussed here. That is fine, as long as they take care how they describe what they have done.

Acknowledgements

Many colleagues and students have helped with the development of these ideas, most recently Brian Logan and Steve Allen. The latter's paper presented at the I3 Spring Days workshop arises out of this project. This work is presented at length in papers in the Birmingham University Cognition and Affect FTP archive. Recent papers are listed in
ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/0-INDEX.html
The project is summarised in:
http://www.cs.bham.ac.uk/~axs/cogaff.html
Our software tools including the SIM_AGENT toolkit (Pop-11 based code and documentation) can be found as part of the Free Poplog FTP site:
ftp://ftp.cs.bham.ac.uk/pub/dist/poplog/freepoplog.html
The SIM_AGENT toolkit is not committed to any particular architecture: rather it supports exploration of a wide range of architectures in single or multiple simulated agents. It also includes teaching materials which can be used to introduce students to some of these ideas through practical experience of designing and modifying simple agents in simulated worlds, including a sheep and sheep-dog scenario, and various kinds of obstacle avoiding and goal seeking agents.

References

[1] L.P. Beaudoin and A. Sloman. A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, and A. Ramsay, editors, *Prospects for Artificial Intelligence*, pages 229–238. IOS Press, Amsterdam, 1993.

[2] L.P. Beaudoin. *Goal processing in autonomous agents*. PhD thesis, School of Computer Science, The University of Birmingham, 1994.

[3] K. Craik. *The Nature of Explanation*. Cambridge University Press, London, New York, 1943.

[4] A. R. Damasio. *Descartes' Error, Emotion Reason and the Human Brain*. Grosset/Putnam Books, 1994.

[5] D.C. Dennett. *Kinds of minds: towards an understanding of consciousness*. Weidenfeld and Nicholson, London, 1996.

[6] J.J Gibson. *The Ecological Approach to Visual Perception*. Lawrence Earlbaum Associates, 1986. (originally published in 1979).

[7] D. Goleman. *Emotional Intelligence: Why It Can Matter More than IQ*. Bloomsbury Publishing, London, 1996.

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

[9] M. L. Minsky. *The Society of Mind*. William Heinemann Ltd., London, 1987.

[10] N.J. Nilsson. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, San Francisco, 1998.

[11] K. Oatley and J.M. Jenkins. *Understanding Emotions*. Blackwell, Oxford, 1996.

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

[13] R. Picard. *Affective Computing*. MIT Press, Cambridge, Mass, London, England, 1997.

[14] H. A. Simon. Motivational and emotional controls of cognition, 1967. Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979.

[15] A. Sloman and M. Croucher. Why robots will have emotions. In *Proc 7th Int. Joint Conference on AI*, pages 197–202, Vancouver, 1981.

[16] A. Sloman and B. Logan. Building cognitively rich agents using the Sim_agent toolkit. *Communications of the Association of Computing Machinery*, 42(3):71–77, March 1999.

[17] A. Sloman. On designing a visual system (Towards a Gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, 1(4):289–337, 1989.

[18] 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.

[19] A. Sloman. Prospects for AI as the general science of intelligence. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, and A. Ramsay, editors, *Prospects for Artificial Intelligence*, pages 1–10. IOS Press, Amsterdam, 1993.

[20] A. Sloman. Explorations in design space. In A.G. Cohn, editor, *Proceedings 11th European Conference on AI, Amsterdam, August 1994*, pages 578–582, Chichester, 1994. John Wiley.

[21] A. Sloman. What sort of control system is able to have a personality? In Robert Trappl and Paolo Petta, editors, *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, pages 166–208. Springer (Lecture Notes in AI), Berlin, 1997.

[22] A. Sloman. Damasio, Descartes, alarms and meta-management. In *Proceedings International Conference on Systems, Man, and Cybernetics* (*SMC98*), pages 2652–7. IEEE, 1998.

[23] A. Sloman. The "semantics" of evolution: Trajectories and trade-offs in design space and niche space. In Helder Coelho, editor, *Progress in Artificial Intelligence, 6th Iberoamerican Conference on AI* (*IBERAMIA*), pages 27–38. Springer, Lecture Notes in Artificial Intelligence, Lisbon, October 1998.

[24] A. Sloman. What sort of architecture is required for a human-like agent? In Michael Wooldridge and Anand Rao, editors, *Foundations of Rational Agency*. Kluwer Academic, Dordrecht, 1999.

[25] A. Sloman. Review of *Affective Computing* by Rosalind Picard, 1997. *The AI Magazine*, 20(1):127–133, 1999.

[26] A. Sloman. Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In Kerstin Dautenhahn, editor, *Human Cognition And Social Agent Technology*, Advances in Consciousness Research. John Benjamins, To appear.

[27] 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.