

Synthetic agents: synthetic minds?

D.N.Davis

School of Computing
Staffordshire University,
Stafford, ST18 0AD, UK

ABSTRACT

There are a plethora of agent definitions. These range from descriptions based on a functional analysis of how agents are used in technology to far more ranging expositions based on different interpretations of the role and objectives of artificial intelligence and cognitive science.

It is possible to establish an ontology within which agents (and their applications) can be characterised, allowing agent definitions to be compared and providing an ontological framework within which the design requirements for synthetic agents, and by extension synthetic minds can be addressed. The contention here is that by developing sufficiently cogent models of (human) minds that are capable of acting as specifications for a synthetic mind, we can not only address the strengths and shortcomings of those models (or theories) through the development of computational models but develop synthetic agents that could be said to exhibit qualities associated with having a (synthetic) mind.

Irrespective of what dialectic we use to analyse the behavioural and cognitive qualities associated with a mind, there are a number of underlying questions that need to be addressed, including: What sort of computational architecture will enable this phenomenon? It is suggested that there is no one architecture and the rest of this paper considers a few alternatives. The discussion is based on experiments with computational agents that address questions related to the architecture, the range of control states, and the behavioural and cognitive capabilities associated with a mind.

1. INTRODUCTION

This paper reports on empirical investigations of complete agent architectures and the related development of a theory of mind. This work is primarily related to the exploration of the possibilities offered by different agent architectures for the modelling of motivational and other control states and the consideration of the computational limitations of a theory of mind. Through moving from theory to design to (any number of plausible) implementation(s), the strengths, inadequacies and oversights of such a model become apparent. As a side-

effect, we hope that the emergent designs are suitable for applications. Furthermore through producing plausible computational models of synthetic agents we may further our understanding of biological, psychological and other possible forms of agents. These are very long-term motivations and we can expect to make slow progress.

When we design and implement complete software agents, particularly if attempting to pursue wide cognitive science goals, there are a number of important questions that repeatedly arise, for example: what are the research issues? what makes a complete architecture? and what are the control issues in a complete architecture? Franklin touches on similar methodological aspects in a recent paper [1].

The rest of this paper presents an overview of (some) related work and some criticisms of them on the basis of their philosophical and psychological plausibility as designs for an synthetic mind. It then considers what kind of agent architectures will possibly cope with the types of motivational and control states identified with a general purpose agent, loosely modelled on the human mind. It is contended that a four level architecture provides an appropriate guiding model. The central theme to this argument is that while implicit representations and related types of processing may be sufficient to model certain types of agent behaviour, these must be somehow integrated with other types of processes if higher level (cognitive) abilities are required. Further architectural and methodological considerations, based on a number of computational experiments and related research, are then discussed. This leads to a discussion on how it may be possible to cleanly integrate the differing categories of behaviours thought necessary for complete (but synthetic) agents. As to be expected, the conclusions are tentative!

2. THE ONTOLOGY OF AGENTS

Artificial Intelligence is a very diverse field, with many (non-exclusive) threads, and agents are used as metaphors for work in most areas. This can lead to confusion and devalue the term agent (as other authors have noted) [2]. This section clarifies the framework of the agent research being highlighted here.

The use of the term agent (in AI and related fields) can be traced back over the last thirty years, with many definitions. A seemingly endless list of agent attributes are possible reflecting these different definitions and the intentions of their creators, and include factors such as intentionality, autonomy reactivity, flexibility, communication, learning, self-actuation etc. [2,3]. What is not obvious from these types of lists, is how these attributes map onto framework for describing a mind. Two different but plausible agent definitions, are:

Definition 1: *An agent is an integrated computational entity with intentionality and some degree of autonomy.*

Definition 2: *An agent is a synthetic entity that enables us to study, at a computational, design or theoretical level, what a mind could and can be.*

The first definition equates to the idea of *weak agents*, i.e. agents as complex (intelligent) information processing systems and is quite open to extension and interpretation. The second definition could be equated to the notion of *strong agents*, i.e. agents as computational cognitive models that explain and/or simulate, to some degree, reported findings and theories in cognitive psychology, or some other study of minds (or life). It is the second of these themes that drives the following discussions.

Architectures for a mind

Brooks along with many other researchers have developed what have become known as *Behaviour Based Architectures* [4,5,6]. These seem to work well for robots that model insect-like or similarly limited task-related behaviours, but seem an inappropriate model for complete agents, which need not only to behave *intelligently* with regard to tasks within their environment, but also demonstrate more expansive qualities, whether it is reasoning about past and future events in their environment, reassessing their role within their known (and possible) environments or more creative behaviours.

An alternative approach (which both pre- and post-dates the reactive behaviour-based argument) allows the mind to be considered as a set of mechanisms capable of supporting a number control states [7,8]. Briefly, it is possible to consider control states through a categorisation of different mental phenomena as discussed in the psychological and philosophical literature. These control states may take several sub-forms as discussed elsewhere [9,10]. This taxonomy, in practice, is very fuzzy and there is much overlap and feedback between these different categories. For example, the desire to maintain a standard based on beliefs about adverse conditions may instantiate certain types of moods. Examples from a non-exhaustive taxonomy of control states would include:

Beliefs which are internal models of the world, possibly inferred from perceptual acts, e.g. *object close, to the left and moving away*. These need not have a rational basis, for example children's belief in the existence of Santa Claus;

Images which are mental constructs related to pictorial and spatial arrangements used, for example, in spatio-temporal reasoning or thinking about the work of visual artists;

Motivators are dispositions to assess situations in a certain way; i.e. a context for reasoning about epistemological events;

Moods are persistent states; they can be viewed as emergent states that pervade the entirety of cognitive processing or a side-effect of other control states. Certain moods favour certain motivators and inhibit others; i.e. they are closely related to predispositions and attitudes.

Specific control states can exist at several levels, requiring different types of processing. Reflexes, for example the reaction to unexpected auditory stimuli, may involve motor (or body) actions and/or deflection of cognitive processing from current tasks and may also influence future high level processes (for example, a motivator to train attention to ignore irrelevant perceptual stimuli when performing certain tasks). There is a problem with this control state theory in that it poses no answers to the idea of the mind as an ongoing characteristic of a living entity of (at least) a certain level of complexity. This and related issues will be addressed in a subsequent section, after a review of a number of empirical investigations of alternative theoretical and design viewpoints.

A further perspective comes from an amalgamation of sources. For example, the OZ project suggests that to build believable agents (which surely complete agents should be), it is necessary to adopt a broad and shallow approach [11]. This extends what Kaelbling calls robustness, that is a "...system should continue to behave plausibly in novel situations and when...impaired" [6]. Other work considers how various models of symbol manipulation, and related experimental evidence, can be combined to produce an acceptable theory of cognition [12].

Combining these approaches entails that work on complete agents requires many mechanisms to be incorporated at a shallow level, enabling investigations of the agents, and therefore the theory, at a holistic level. For example, we can provide models of the multiple perceptual paths (to be found in biological agents) but limit the computational detail associated with the processes. The initial work described here focused on the inducement of goals from internal states and external events, with some depth to the directly supporting processes but shallow (or minimal) implementations of other processes thought necessary for a complete agent.

3. THE FOUR LAYERS OF MIND

Exploring design space for (cognitively) complete agents is a lengthy (currently impossible?) task. Here a number of design and computational experiments based on a four layer model of the mind are considered. These are based on a developing architecture (and theory) of mind arising from the work of Sloman and others [8,9,10,13,14]. Figure 1 presents a highly stylised view of an architecture that aims to support the types of (internal control) behaviours referenced in the preceding

section. This model is quite general, and the effect of altering the relative size and importance of the layers is an open issue. Unlike some other theories (e.g. Newell), this presents a broader picture of the mind, with high level (*cognitive*) and low level (*conative*) processes coexisting and interacting in a holistic manner. In effect, goal processing, planning, decision making and other cognitive processes are not purely abstract but exist in relation to other automatic (perhaps unconscious) processes. They are, in effect, embodied within the context of their interactions with their underlying processes, and the agent's relationship(s) with its environment.

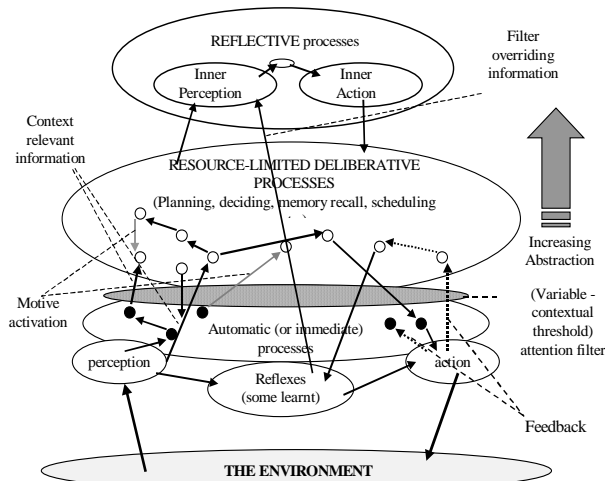


Figure 1. A stylised representation of a plausible architecture for a complete agent: the "initial" extended Sloman model.

At the base level exist reflexes, which are very fast (inherently parallel) pre-attentive (or immediate) processes that allow a direct response to sensory events. This does not preclude them from provoking processes (or being modified in some other way) at a more abstract level. Other automatic processes are similarly pre-attentive, but *necessitate* the generation of conscious control states to achieve their goals. The deliberative layer represents those processes typically studied in thinking, human problem solving etc., plus other processes related to the management of low level actions. The reflective processes serve to monitor cognitive behaviour or control it in some other way. It is suggested that learning mechanisms (of one form or another) permeate the entire architecture.

The attention filter, which mirrors results from the psychological study of attention, exists to protect the more resource limited (and serial or limited parallel) deliberative processes from unwanted and unnecessary interruptions. The attention filter may be context loaded and/or respond to signal strength. Some low level actions (for example, life threatening events causing fast responses at the pre-attentive level or more sublime events such as the cocktail party phenomena) can effectively override the filter on their signal strength or context [15]. In other situations the filter accepts information that is

related to ongoing (or stalled) deliberative processes. At times, this filter would provide less of barrier, mirroring low processing activity at the deliberative level or a loss of interest in the focus of the ongoing internal activities.

In moving from this (admittedly sketchy) theory towards a computational entity, the design stage offers the researcher a plethora of viable alternatives, and a number of design dilemmas. A biologically plausible design for the lower levels of the architecture (e.g. neural net based a-life forms) may not support the processes associated with the deliberative layers (for example, reasoning about concepts).

The main initial concern was in connection with the concepts and processes associated with motivation, in particular: how motives are generated and managed. This required an investigation of the internal structure of motives, how motivators are processed including how new ones attract attention, how their importance is assessed and how they are assessed and utilised. Further more expansive questions relate to how conflicts are dealt with, how resource-limited processes are managed (i.e. how attention is controlled), how plans are formed or selected, and executed, how new events can affect old motives, and so on. Some of the processes involving motivation can lead to states that are at least partly analogous to some of the states described as *emotional* in humans. Earlier computational work on this theoretical architecture was very limited, and failed to integrate the different layers to the theoretical design [16]. In order to further investigate the theory, a number of design (and implementation) experiments were performed.

Design Experiment One

This experiment considered how the deliberative layer processes (which were designed as integrated and inseparable) could monitor an environment and manage a set of reactive agents moving around and performing tasks in that synthetic world. We made use of an ongoing research scenario, which allows us the benefit of comparing the different classes of agents in similar environments. This scenario makes use of a two dimensional world within an agent toolkit. The simulated environment consists of one or more rooms, and is potentially hazardous containing dynamic and static agents and objects.

The deliberative processes included some rudimentary planning, a two valued Boolean model of beliefs about the (control) state of agents in the environment, and a stylised iconic model of the environment. The reactive behaviours include goal-based behaviours such as finding and moving trapped simple agents to free spaces. The reflexive behaviours used were simple movement behaviours (stop, turn etc.) to avoid obstacles.

Control states (as discussed above) can be implicit or distributed among coexisting processes and memory structures. For example, the behaviours (reflexes) associated with the lowest architectural level are implemented in such a manner. Some control states, for example certain types of goals, can be

explicitly represented. Goals may be nested; for example the top-level goal associated with one motivational state (*free trapped agent*) is to find a free space for a trapped agent; this may require sub-goals such as locate and move to the trapped agent, and then subsequently move to the free space, which, of course, may no longer be free.

For the reactive-deliberative agent combination, our main concern was with the gener-activation and management of goals [10,14]. We made use of a deliberative goal database, and an attention filter with thresholds associated with the insistence value and the deliberative contexts of the adopted goal structure. If further (reactive) goal creating states are generated (in subsequent time intervals) and their insistence value is greater than the attention filter threshold, then goals can be created and added to the goal database; potential goals not meeting this criteria are simply deleted. An executive process, on noticing goals in the goal database, ranks these goals in terms of their intensity (a combination of importance and urgency). The most intense goal is considered to be a candidate for current deliberative processing if ranked more important than the current focus of attention.

This approach of tightly integrated deliberative processes enabled us to more fully consider the nature of motivational processes, and further extend the characteristics and structures used for representing and handling goals. It unfortunately raised a number of challenges related to the integration of decision mechanisms across the different layers of the model. In particular the adopted feedback mechanisms which should give rise to learning processes were inappropriate for these purposes (at a design level). Furthermore, the belief model was inappropriate for generating appropriately parameterised goals; a more expansive memory model was needed with a signal strength associated to items stored in memory. It was also difficult to incorporate the use of the most abstract reflective processes, and for the agents to merge compatible goals across and within different layers.

Design Experiment Two

This experiment, again driven by the four layer model, made use of a more distributed design (see figure 2), in an attempt to overcome some of the difficulties encountered in experiment one. The domain was the same but the agents' capabilities, at all levels, were extended. The fundamental idea was that each architectural layer could be modelled in turn using agents of increasing sophistication, and that for any class of agent different sub-types, with complementary capabilities, can exist. The theory was also extended to allow deliberative contexts to separate the reflective and deliberative layers.

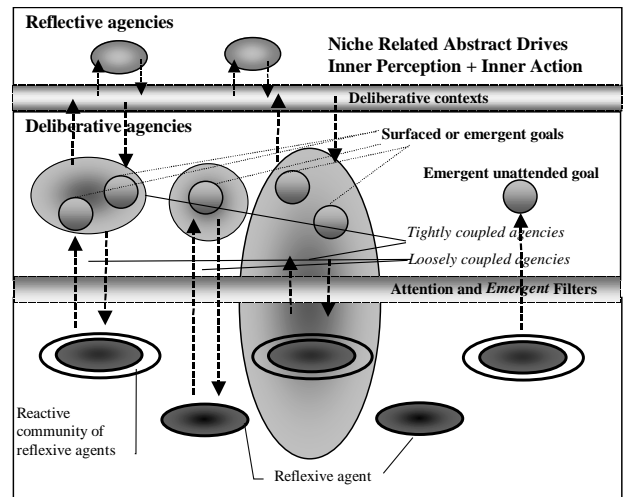


Figure 2. Abstract architecture of the society of agencies approach to complete agent design.

The base level (reflexive) agency (a mobile plinth, modelling the perception-reflex-action) is allowed some degree of autonomy and must navigate the environment, and find its energy source, without resource to persistent representations or higher level reactive or deliberative processes. Slightly more sophisticated (reactive) agents can subsume these mechanisms and include other pre-attentive processes enabling more flexible behaviours (and control mechanisms) to allow them to perform a wider range of possibly more complex tasks; but again without persistent representation or access to the deliberative layers. This reactive class of agent, for which a behaviour-based design seemed appropriate, provides a computational platform for the resource-limited management processes and allows the integration of reactive and deliberative processes to be investigated.

Other agents build upon these simpler agents to allow implementations that more closely represent the fullness of the architecture hinted at in figures 1 and 2. The next class of agent maintains a persistent representation of the environment, at the deliberative level, that allows beliefs and memories. These agents can attend to a limited number of goals and make use of the attention threshold. The design of the memory mechanisms allows inconsistent beliefs, with fuzzy valued strengths associated with memory statements. As new information is added to memory, it reinforces similar information but weakens older disagreeing information; sufficiently weak memories are deleted. The memory strengths in turn can then be associated with (pro-active) goals that are generated on the basis of beliefs and memory. This requires the deliberative agents to reason with memories and beliefs about other agents' behaviour over time, and extends the recursive agent modelling capabilities. Some limited (implicit) forms of attitudes and standards are associated with the generation of pro-active goals; for example, an altruistic (deliberative) drive to care for hapless agents in

the environment. A further attitude is related to some deliberative agents' desire to inhabit environmental space uninhabited by other deliberative agents; causing the generation of pro-active goals, based upon context-related perceptual information.

A number of different types of decision-making processes are used in the different classes of agent. This particular approach to modelling agents, with competing and co-operative behaviour modules sometimes acting independently of each other, can cause problems in that we need to provide some means of deciding between conflicting potential actions (at the various levels in the architecture). Sometimes different actions are not in conflict and parallel or sequential combinations can be adopted. For example, a reflexive agent cannot turn left and right at the same time, but could turn left and stop, or start and turn right. For reactive and reflective agents with a number of levels of processing a more sophisticated approach is required. In short, an agent prefers to make decisions, using subsumptive mechanisms, at the lowest possible level; a reactive agent will act like a reflexive agent unless it requires reactive processing, and similarly a deliberative agent will act as a reactive agent unless deliberative behaviours are required. This ensures that the computationally more expensive higher level processes are used only when and where necessary. It should be noted that an agent will not necessarily prefer behaviours related to higher level processes over those generated at a current level.

This experiment provided a more satisfactory mapping of the theory than experiment one, but unfortunately external factors required the curtailment of this experiment. There were still a number of inadequacies. Again learning mechanisms for updating (and optimising) behaviours proved to be troublesome to incorporate. At the design level, it is possible to see how different categories of learning segue well with certain capabilities: for example, Q-learning for the (subsumptive) decision mechanisms in the reflexive and the reactive-reflexive agents; and again perhaps inductive learning over goals and belief models at the deliberative level. But in practice, this proved to be difficult to implement, without considerably changing the design (and extending the theory in a number of ways). It was also felt that the niche roles available for these agents, in the simulation, were too limited and a more expansive domain would allow us to address the theory more critically.

5. REVIEW OF PROGRESS

So what have these (and other) experiments shown? The group work, initiated by Sloman, has been advanced through a series of implementations based on philosophical and psychological foundations. This is difficult and slow going - but who said developing an artificial or synthetic mind was easy? More importantly, some slight progress has been made at a design and theoretical level.

It can be argued that a number of implementations have demonstrated some "emotional" states; in particular emergent

perturbant states. It is due to the competition between different (pre-cognitive and cognitive) activities that cognitive effects such as emotional perturbation can be explained. Disruptive emotions arise from conflicts between antithetic behaviours associated with these control states. The opportunistic or deliberate satisfaction of (some) goals or other motivators, should enable a more harmonious emotion. A cognitive model, that truly claims to model a human mind, would have to cope with multiply active, not necessarily co-operative, states and ensure that any emergent or designed patterns of behaviour in these circumstances is believable.

Nothing in the work, so far, has really undermined the four layer principle to a theory of mind. Furthermore, the control state approach is still valid, given that it is not an exhaustive description of what could be occurring within a mind and that these control states need not be symbolic in nature. There are, however, inadequacies with this loose theory. For example consider how the most abstract layer relates to many of the control states. It is suggested that personality traits may operate at this level, and therefore permeate the rest of the architecture. If the mind is an ongoing characteristic of a living entity of (at least) a certain level of complexity and a mind is capable of moving through all these other control states, from where do the control patterns that stabilise a personality emanate? Personality traits can be seen as control states that affect the reflective processes and influence the types of deliberation (e.g. decision making) that cause cognitive and animated behaviour. Personality then becomes an emergent property of the cognitive architecture and its disposition to concentrate on certain tasks and favour specific control states. Alternatively, perhaps it is solely an emergent property of (possibly low-level) drives. Or perhaps it is a combination of these and other more abstract reflective processes, and that these latter processes are, in part, responsible for the persistent attunements to global niche space. Moods can therefore arise from the interaction of current temporally-global drives and temporally-local low-level drives to reflect the current focus of the deliberative processing as perceived by the reflective layer. Temporally-global drives are those associated with the agent's overall purpose (i.e. its niche role), while temporally-local drives are related to ephemeral states (or events) within the agent's environment (or itself). The (high-level) niche-seeking drives together with the more orthodox control states, in a synergistic arrangement of deliberative and reactive processes with the low-level drives, are what could bind the theoretical model together and allow a synthetic agent to become complete and exhibit a (non-shallow) personality.

At a design (and implementation) level, experiments one and two quickly reached some of the problems highlighted by Franklin [1]. How do we integrate learning in these multi-level computational models. The problem is analogous to the experimental work of Lashley; where does memory of learnt behaviours in biological agents (rats in his case) reside [17]? The design and attempted implementation of a centralised

memory and also learning mechanisms seems analogous to Lashley's (fruitless) search for the engram (i.e. the neurophysiological basis of memory). No one single process (however connected) should account for these mechanisms at a theoretical, design or computational level. Any attempt to do so is almost bound to arrive at the problems highlighted here and in Franklin's methodological approach to designing complete agents. In order to develop the work further, we need: further analysis of the requirements for complete agents; extend both the breadth and the depth of the deployed architectures, including the integration of various theories of learning; and explore the philosophical, psychological and engineering objectives of AI.

Finally, I want to consider a thought experiment in potential designs for a distributed *emergent* architecture, of use in agents for game playing, decision-support and creative processes [18]. Not only can we specify the initial capabilities and behaviours of the agents, but also the initial conditions for agents to collaborate, whether at the same level or across different levels. This will build on the ideas hinted at in figure 2, allowing heterogeneous communities of agents that mirror multiple cognitive (and conative) activities. While this may lead to co-operative and competitive cliques of emergent processing, there is the possibility that these communities may overwhelm the processing of the overall agent, leading to undesired control states. However within the theory there are mechanisms, for example processes emanating from the reflective layer, that could effectively coerce these groups towards more directed or effective patterns of interaction. This type of combination may allow a reflective (temporally-global) navigation of deliberative processes and control state related communities of (reactive and reflexive) agents, and achieve a more substantial demonstration of the possibilities associated with complete agents. Learning mechanisms in this scenario are agent frameworks within which certain classes of agencies, and perhaps communities of agents, have their mechanisms or processes re-calibrated to suit changing environmental or niche-role pressures, including the epistemological basis of those processes. Perhaps this proposed marriage of a-life and agents will shed some well needed light on how to integrate different categories of learning across agent architectures and provide further analytical depth on the nature of other control states within the context of complete agents.

6. ACKNOWLEDGEMENTS

This research was and is funded through a number of sources: University of Birmingham vice-chancellor's grant; EPSRC travel grants; Staffordshire University Research Initiative.

7. REFERENCES

1. S.P. Franklin, Autonomous Agents as Embodied AI, *Cybernetics and systems*, Vol. 28, No. 6, 1997, pp. 499-520.
2. S.P. Franklin & A.G. Graesser. Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents, In:

- Intelligent Agents III*, J.P. Muller, M.J. Wooldridge & N.R. Jennings (Eds.), Springer-Verlag, Heidelberg, 1996.
3. M. Wooldridge & N.R. Jennings (Eds), *Intelligent Agents*. Springer-Verlag, 1994.
4. R.A. Brooks, Intelligence without representation, *Artificial Intelligence*, 47:139-159, 1991.
5. P. Agre & D. Chapman. PENGI: An implementation of a theory of activity. *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, 1987.
6. L.P. Kaebling, An Architecture for Intelligent Reactive Systems, *Readings in Planning*, Morgan Kaufmann, 1989.
7. H.A. Simon, Motivational and emotional controls of cognition, *Models of Thought*, Yale University Press, 1979.
8. 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.
9. D.N. Davis, Architectures for Motivated Agents. *Journal Of Computational Intelligence (Under Review)*, 1998.
10. A. Sloman, L. Beaudoin & I. Wright, Computational modeling of motive-management processes, In: *Proceedings of the Conference of the International Society for Research in Emotions*, N. Frijda (Ed.), ISRE, 1995.
11. J. Bates, A.B. Loyall & W.S. Reilly, Broad agents, *SIGART BULLETIN*, Vol. 2, No. 4, 1991.
12. A. Newell, *Unified Theories of Cognition*, Harvard University Press, 1990.
13. A. Sloman. What sort of architecture is required for a human-like agent? *Cognitive Modeling Workshop, AAAI96*, Oregon, 1996.
14. D.N. Davis, Reactive and motivational agents: towards a collective minder. In: *Intelligent Agents III*, J.P. Muller, M.J. Wooldridge & N.R. Jennings (Eds.), Springer-Verlag, 1996.
15. E.C. Cherry, *On human communication: A review, a summary and a criticism*, MIT Press, 1957.
16. L.P. Beaudoin, *Goal Processing in Autonomous Agents*, Ph.D. Thesis, School of Computer Science, University Of Birmingham, 1994.
17. K. Lashley, *Brain Mechanisms and Intelligence*, New York, 1963.
18. D.N. Davis & B. Berbank-Green, *Game Playing Agents*, Unpublished, 1998.