

## **ARCHITECTURES FOR COGNITIVE AND A-LIFE AGENTS**

Darryl N. Davis

Neural, Emergent and Agent Technology Research Group

Department of Computer Science

University of Hull

Kingston-upon-Hull

England

HU6 7RX

Email: D.N.Davis@dcs.hull.ac.uk

Tel: +44-1482-466469

Fax: +44-1482-466666

Web: **<http://www2.dcs.hull.ac.uk/NEAT/dnd/index.htm>**

Dr Davis is a lecturer in the Department of Computer Science at the University of Hull. He has a B.Sc. in Experimental Psychology, M.Sc. in Knowledge Base Systems and Ph.D. in Diagnostic and Investigative Medicine. He has worked in human visual perception and has over 14 years experience in artificial intelligence systems. These have been successfully applied to classification and computer vision problems in business, medicine and geology. He has research interests in cognitive science (in particular cognition, motivation and emotion), agent technology as a metaphor for the mind and as a vehicle for domain applications, the application of computational intelligence to medical domains and machine vision. He has published widely in all these fields.

## **ABSTRACT**

In this chapter research into the nature of drives and motivations in computational agents is visited from a perspective drawing on artificial life and cognitive science. The background to this research is summarized in terms of the possibility of developing artificial minds. A particular cognitive architecture is described in terms of control states. Steps towards producing an implementation of this architecture are described by means of experimentation into the nature of specific control states. The design of a simple a-life architecture for a predator-prey scenario is described using a transition state diagram. This architecture is then used as a platform with which to develop an agent with drives. This second architecture is used to develop an agent with explicit motivations. The discussion of these (and other) experiments shows how these simple architectures can help to provide some answers to difficult research questions in cognitive science.

## **INTRODUCTION**

This chapter explores the use of artificial life metaphors in providing a context for cognition and affect and provides a discourse on the possibilities offered by artificial life for modeling aspects of cognition, specifically drives and motivations. It is suggested that these phenomena will have an important part to play in the future of agent software engineering as a means for bridging the conceptual gap between agent autonomy and agent benevolence. In this research the term agent is placed in a broader perspective than the current trend in agent-oriented software (Ciancarini & Wooldridge, 2001; Ferber, 1999), relating to the use of agent in, for example, cognitive science (Wilson & Keil, 1999) and philosophical works (Ryle, 1949). A relatively simple a-life scenario is used to demonstrate the relation between theoretical concepts in the research areas of

cognition and affect and computational architectures. The paper highlights a design methodology that allows incremental agent sophistication to be achieved. These explorations and designs make use of the SIMAGENT toolkit – an agent-based toolkit that runs in the Poplog environment.

This chapter describes a framework used to investigate cognitive agent architectures. The framework is a synthesis of a number of approaches and aims to provide a means to map abstract theory to well-defined implementation. Influences include the control state approach to cognition (Simon, 1967; Sloman, 1993), the niche and design space perspective on mind (Davis, 2001b), the broad but shallow perspective to agent development of the CMU OZ project (Bates et al, 1991) and the architectural parsimony of Hayes-Roth (1993). It makes use of logic and finite models (including cellular automata) to specify process and behaviour in an agent.

## **RESEARCH OBJECTIVES**

This research pursues the following three major objectives:

- A synthesis of concepts for studying cognitive agents based on an analysis and investigation of how different perspectives on emotion, motivation and autonomy map onto computational frameworks.
- A framework for developing rules on how artificial life and emergent behaviours can be combined with the more abstract decision making processes associated with cognitive agents.
- Insights into the nature of heterogeneous mechanisms across different processing, representational and knowledge levels, and their explanatory role in describing mental phenomena.

There are controversies about the terminology used in this area of cognitive science research. Motivation, drive, goal and emotion are used to refer to and mean a number of different things. There is no universal definition of these terms across (or even within) the fields of philosophy, psychology, cognitive science and artificial intelligence. Some (Ortony et al, 1988; Frijda, 1986) consider emotions to be a cognition centred set of phenomena, while others (Frankel & Ray, 2001; Chapman, 1996) consider these terms to be centred on low-level (neuro-/physiological) control processes that affect cognition. These arguments are not addressed in depth here – research on emotion is discussed elsewhere (Davis 2001a). However such terms will be defined in the following sections as required and used to refer to specific phenomena pertinent to this work.

Other major controversies relate to the question of whether a-life entities can be said to be alive or mere process simulacra? (Boden, 1996). These controversies are akin to the arguments about the nature of intelligence and how best to investigate, design and implement it (Brooks, 1991; VanLehn, 1991; Wilson & Keil, 1999). A fundamental issue is the difference between Cartesian and biological based cognitive science (Wheeler, 1997). From a Cartesian point of view the rules governing physical bodies and those affecting psychological phenomena are distinct and non-interacting – mind is separate to body. From the biological-based view of cognitive science such a distinction cannot be made. The question then arises that if by using synthetic agents in synthetic environments are we in fact studying synthetic minds which are as alien to those associated with biological entities as silicon is to carbon (Davis, 1998)?

## **BACKGROUND**

The research described here makes extensive use of computational design and experimentation. In moving from theoretical perspectives to alternative designs and then to possible computational models, some research questions are answered without requiring implementation. This is termed the design stance. What is typical of this approach is that in moving through these different levels, flaws in our understanding are revealed and insights into the nature of our research questions compounded by further questions. Beyond the designs, alternative implementations are possible. Often it is the case that what might seem trivial at the theoretical level is complex at the implementation level, even in shallow experiments. This is in part analogous to the challenges presented by agents to orthodox software engineering. Agent technology is advancing fast with an associated increase in the software engineering perspectives appropriate to building agent systems (Ciancarini & Wooldridge, 2001). For some well defined applications, for example e-commerce (Luo et al, 2002), it is possible to combine knowledge engineering principles (Scheiber et al, 1993) with toolkits (Nwana et al, 1999) that extend object-oriented software engineering techniques to encompass aspects of agent programming. However for more investigative experimentation with ill-defined concepts, as investigated here, there is a shortage of easily manipulable toolkits. The agent toolkits that do exist for such work, for example SIMAGENT (Sloman & Logan, 1998), require advanced programming capabilities and prolonged and frequent use to achieve results acceptable to the research community. However they do provide a high-level departure point for experimentation into agent architectures and parameterisation.

A number of problem domains and test-beds have been used for experimental purposes. The choice of domain and test-bed is determined by the research questions to be addressed. If the

major concern is about design options and how to resolve low-level behaviours or the study of collective emergent behaviour in simple agent architectures, use is made of a number of variations on TileWorld (Hanks et al, 1993). In studying motivation and its relation to collective activity we can use TileWorld, an artificial robot controlled factory (Davis, 1997; Davis, 2001b) or RoboCup (Nunes, 2001). In studying the mechanisms associated with cognition and affect, there are no conclusive assumptionless analyses that invalidate this approach. On the contrary, by using simulated worlds greater progress can be made in determining which of the research issues can be tackled this way, which cannot and why that is the case. The research into computational architectures for motivation, emotion and autonomy does not require the use of robots. The behaviours, processes and information structures of interest in this research can be studied without compromise in synthetic environments. The assumption underpinning this research perspective is that these mechanisms in a synthetic mind should be inherently similar across domains, irrespective of the nature of the situated action.

Here a simple predator-prey scenario is used as an exemplar. This environment is a continuous spaced two dimensional world of objects and agents. Objects are not mobile, exist for a specified period and can be used as an energy source by some agents but act as an obstacle to others. Agents move around in this world consuming energy. Agents can sense objects and other agents in their environment if within a specified range. This form of sensing is a simulated perceptual mechanism provided by the agent toolkit. Agents can reclaim energy by consuming specific classes of objects and if predatory, specific agent classes. Agents breed in certain conditions, the offspring being some combination of its two parents. Agents, in moving, deposit an energy trail of agent-specific RGB-valued pixels in a colour image. Agents with appropriate capabilities can

sample and sense this colour image to provide clues to the whereabouts and nature of other agents. This use of machine vision in the agents allows us to deepen the perceptual capabilities of our agent architectures and permits explorations into sensory fusion when combined with the in-built sensing mechanism of the toolkit.

## **DESIGNS FOR MIND**

CogAff (Sloman, 2001) is a three-column architecture of perception, central processing and action. It provides the means to investigate the modeling of mental phenomena. It consists of three-layers: reactive mechanisms which are (sometimes ballistic) behavioural responses to environmental and internal states; deliberative reasoning which requires explicit manipulation of some form of cognitive structure (for example motivational constructs); and meta-management mechanisms (reflective processes) which monitor the agent's internal state, processes and its ongoing stimulus-response arcs.

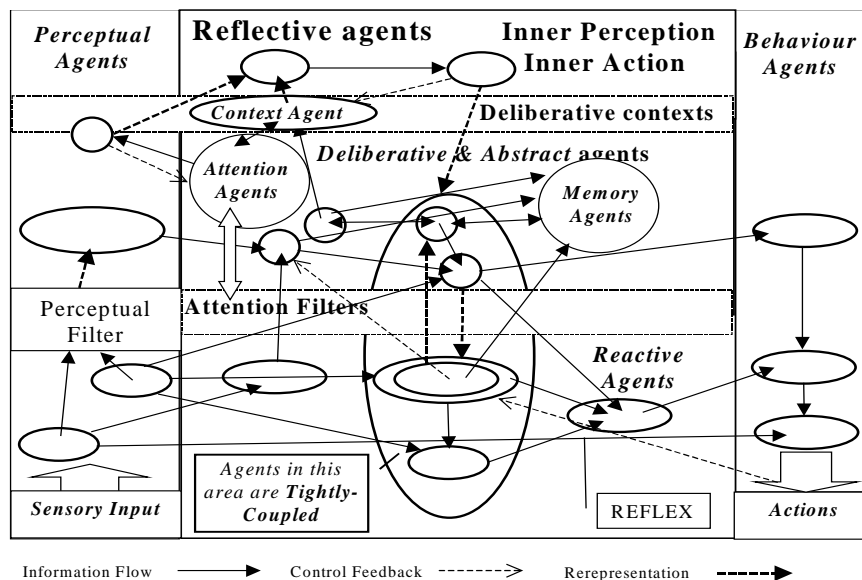
This very abstract architecture accepts perceptual information and outputs action in the environment. Alarms provide feedback from action and other intermediate stages across the three layers and columns. Personae are associated with the reflective level. As different personae become active, the control state of the entire agent changes. This affects the interpretation of incoming information and the actions of the agent in any given external situation. For example if modelling the type of visual perception described by Gibson (1986), sensory data when perceived is given affordances and valences. An environmental event in the path of an agent will be perceived according to the nature of the current control states. Certain events may be control state neutral, others may exist in antithesis to further control states. Consider a second agent in danger and placed in front of such an agent. If the perceiving agent has adopted a selfish personae, this

external event will be given a negative valence and afforded avoidance behaviours. If the perceiving agent has adopted a selfless personae, this external event may be given a positive valence and afforded rescue behaviours. Some events may be completely irrelevant to the agent and are ignored or afforded neutral obstacle-avoidance behaviours. Other events may be outside of an agent's immediate repertoire of behaviours but deemed important enough to activate further deliberative processing (for example motivations).

From the control state perspective, an agent's immediate repertoire of instantiated drives and behaviours can be considered to be transient control states. Motivations are more persistent (but ultimately temporary) control states that may or may not segue well with the agent's current set of personae. Conflicts between control states can therefore emerge. An intelligent and truly adaptable and autonomous agent should be able to resolve these conflicts through internal changes. This type of phenomena is associated with emotions in human and similar affective states can be found in synthetic agents (Sloman & Croucher, 1987). For example, an agent in need of energy but not sensing any may activate a motivator or goal to find an energy source. The behaviours required to realise this motivation may exist within the agent's overall repertoire of behaviours but be at odds with the desire or intention to explore or avoid certain aspects of the agent's environment. This state of affairs can be resolved using control state concepts, and the motivator management information used in resolving this conflict stored with the motivator. This cannot be said for most structures used for representing goals. It is possible that most agents need to resolve these sorts of conflicts and the designers do not realise that in producing such agents, they risk compromising the concept of an autonomy that underpins the agent.

At the first ATAL workshop, autonomy was defined as one of four foundations for a weak notion of agency (Wooldridge & Jennings, 1995). Autonomy was defined as operating “...*without the direct intervention of humans or others, and have some kind of control over their actions and internal state*”. Castelfranchi (1995) categorizes and discusses the various types of autonomy that a cognitive agent (or robot) can demonstrate. In particular, a distinction is drawn between belief and goal autonomy in the “Double Filter” Autonomous Architecture. If an agent is to be autonomous, then it must set and have control over its goals. External agents, whether biological or computational, should not have direct control over the setting of an agent’s goals, but can only influence these through the provision of information that affects an agent’s belief set. The management of belief sets is described in terms of rationality (logic based reasoning) and credibility of incoming information (sensing). An agent can only be autonomous within certain bounds. Agent autonomy can be constrained by design specifications that require the agents to pursue certain goals on behalf of their designers (*agent benevolence*). In sophisticated applications with agents dealing in uncertain information and multiple goals, it would seem reasonable to expect conflicts to arise from this compromise between autonomy and the pursuit of possibly incompatible goals. The more encompassing control information associated with motivations in this work may provide agent designers with a tool by which such conflicts can be resolved. Our designs for intelligent systems are such that the resulting computational system should be capable of monitoring itself and catching perturbant behaviors before they become disruptive. Furthermore an intelligent system should also be capable of recognizing and harnessing beneficial emergent behaviors.

There is an argument running in the a-life approach to cognition that this type of cognitive architecture with its various modules and components cannot be designed but needs to be evolved (Husbands et al, 1993). The central thesis to that argument is that there are so many modules with so many interconnections that it is beyond the capabilities of the human mind to design and implement. We counter this argument by adopting a distributed approach to mind. This approach has its computational roots in the early work on agents and distributed blackboard systems (Erman & Lesser, 1975). A society of agents with specific and differing capacities and capabilities can be designed and implemented. The interactions within this society of these agents then models the interconnections thought impossible to model explicitly. The thesis central to this counter-argument is that the collective and emergent processing of a suitable designed society of specifically engineered agents will give the required overall capabilities. However it is important that such an agent community has the capability to recognise and manage its emergent behaviours. This is the line being pursued in the research described here and sketched in figure 1.



*Figure 1. A societal approach to cognition based on the CogAff architecture*

We aim to demonstrate that sophisticated cognitive behaviour can arise from the interaction of suitably designed agents in much the same way as demonstrated for socio-economics in Sugarscape (Epstein & Axtell, 1996). The remainder of this chapter will show how to pursue a line of research into these phenomena using computational tools guided by the principles of incremental complexity and architectural parsimony. This experimentation proceeds from relatively simple, easily realised architectures through to quite complex designs.

### **DRIVES AND A-LIFE AGENTS**

Drives are low-level, ecological, physiological and typically pre-conscious. They provide the basis for an agent's behaviour in the world, are periodic but short-lived and are defined in terms of resources essential for an agent. Such activities for information agents include the need to gather resources and propagate information to associates in their society. In biological agents such drives include thirst, hunger, and reproduction. Thresholds for the onset and satiation of such drives are variable and dependent upon processes internal to an agent and external factors arising in the agent's environment. Such drives can be modeled relatively easily in computational agents. These simple "a-life" agents then provide a theoretical and computational platform with which to explore more interesting phenomena.

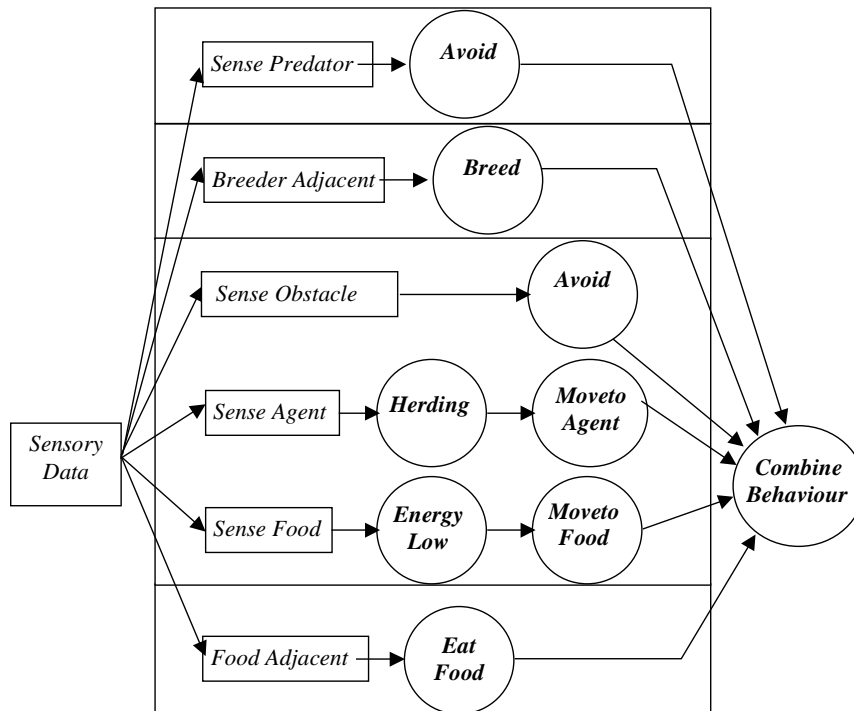


Figure 2. State Transition Diagram for Drives in the Simplest Prey Agent (Agent0)

Consider a base level agent (Agent0) with no explicit motivations but three implicit drives: the need to eat, escape predators and reproduce. Behaviours associated with these drives can be modeled using state transition diagrams as in figure 2, described using propositional calculus and implemented as an action selection architecture (Agre & Chapman, 1987) in SIMAGENT. These agents move around an environment populated with food items and other agents (we have experimented with hundreds of such agents in any one environment), some of which are predatory. The agent can sense items within a specified radius of itself, and discriminate between food, non-predatory and predatory agents. This agent is given an initial energy level and must locate food items to maintain its energy. Movement causes energy to be consumed in proportion to its velocity. When the energy level dips below a specified threshold the agent changes state (it becomes hungry) and must sense and find food. If this energy level reaches zero the agent dies.

The agent can be in a herding or non-herding state. When in a non-herding state all non-predatory agents are effectively mobile obstacles to be avoided. Irrespective of state, predatory agents are mobile threats and need to be avoided. If their energy level is sufficiently high and another agent is adjacent, agents can reproduce. The resultant agent drains a percentage of the energy of both parents and is initialized with a mix of its parent's internal information. The Eat-Food reflex (bottom of figure 2) can be activated regardless of any other potential action or internal state given that food is adjacent to the agent. Each node in the state transition diagram can be formalized using propositional logic. These formalisms then map onto object-oriented methods and production rules within the SIMAGENT toolkit.

Velocity and direction of an agent are governed by a two-part vector representing velocity in vertical and horizontal directions. A change in direction, towards or away from any other object or agent, is expressed as a change to this vector. The actuation of any behaviour in an agent maps onto a vector change impulse. Where more than one behaviour is actuated, the combined impulses are summed in the "Combine Behaviour" node of figure 2. This simple (non-)decision mechanism results in an agent responding to all inputs. Hence a hungry agent may move towards a sensed predator if sufficient food items are similarly located. This agent makes for a useful plinth (left hand side of figure 5) with which to experiment further.

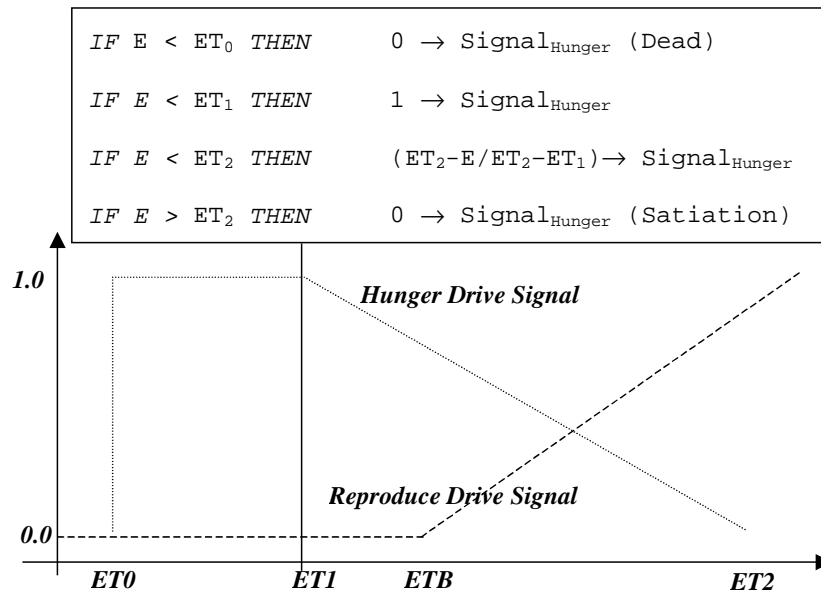
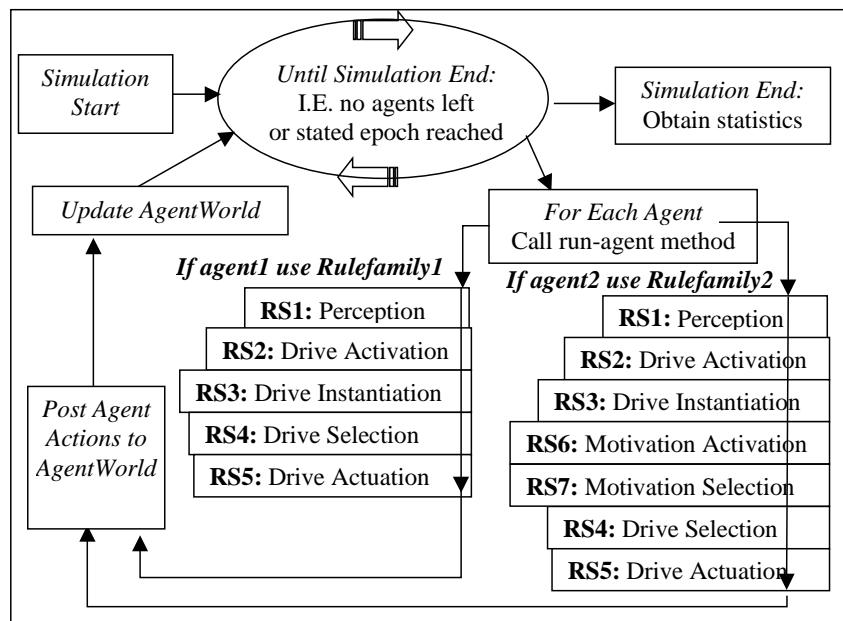


Figure 3. Model for internal energy based drives. The four energy thresholds have been subject to experimentation and could be optimised using for example genetic operators.

The next class of agent (Agent1) extends the model associated with the drives of the base plinth. In Agent0 drives are either on or off. Drives in Agent1 makes use of signal strength based on two factors – an internal measure (energy) and an external measure (distance from some object or agent). The signal strength based on internal energy (used in the food and reproduce drive) is defined as shown in figure 3. The signal strength associated with the Hunger drive makes use of the current energy level (E) and three thresholds as given in figure 3. The thresholds can be optimized through experimentation and analysis and also subject to change through agent reproduction. A similar model is used for the drives associated with the agent’s perception of its external environment. Again these models readily map onto the object-oriented methods and production rule capabilities of the agent toolkit.

Specific capabilities (for example perception, drive activation, drive selection and behaviour combination) are described using propositional calculus and then bundled together as rules in

rulesets (see figure 4). For example one ruleset (RS1) deals with agent perception, while ruleset RS2 determines which of the agent drives can be currently active. A further ruleset (RS3) instantiates active drives with specific values for signal strength and target locations. A fourth ruleset (RS4) selects compatible drives, and the final ruleset (RS5) maps these drives onto agent actions which are posted to the world update methods of the agent toolkit. Rulesets are bundled together as rulesystems (e.g. RuleFamily1 for Agent1) which define the information processing ontology for any particular class of agent. Alternative rulesets for any specific behaviour allow the agent to switch between alternative processing modes. Rulesets can be shared across rulesystems and more complex agents build on simpler agents by extending the rulesystem with further rulesets (e.g. RuleFamily2 in figure 4).



*Figure 4. AgentWorld of Methods, Rule-Sets and Rule-Families.*

Agent1 is a reactive architecture, responding to both internal and external pressures. It can make use of the same decision making mechanism as Agent0 but with the impulses (i.e. vector

changes) weighted according to their signal strength. An alternative decision making mechanism is to select that impulse with the highest signal as a primary goal and then any other impulse that does not conflict with this behaviour. For this test harness all agent behaviours result in a movement in two-dimensions. Hence a conflict in behaviour is simply a vector that distracts from the direction vector of the primary goal. Again there are experimentally derived parameters that define a maximum deviation to the nominal direction vector. Agent1 displays drives, implicit goals (the nominal direction vector of the drive with strongest signal) but no explicit motivation. Any affective qualities are an emergent property of the agent and its environment. Maundering, the switching between two goals or behaviours, is one example. Consider a simple experiment where a predominantly static predatory agent is flanked by two food items of benefit to a further agent. While the energy level of the non-predatory agent is high it will move away from the predatory agent. As the energy level drops, the signal strength to move towards a food item increases. At some point the signal strength associated with moving towards the food becomes sufficient that it is adopted as the primary goal. The agent therefore moves towards one or both of the energy sources (these are not antithetic behaviours) and hence the predatory agent. The behaviour to move away from a predator is then given a higher signal strength and the agent moves in its original direction away from the predator and hence the food. Typically the behaviour of the agent fluctuates until it runs out of energy or is consumed by the predator. If vector summation is used instead, the agent oscillates around its original position but again gradually moves towards the predator with the same result. This agent is in effect caught in a cycle of goal conflicts and maunders between acting on the flee-predator and seek-energy drives. The agent can possibly reach the energy items through planning. A reactive planner with precompiled plans may not necessarily produce a non-direct path to the required items and may

therefore result in the same behaviour as described above but at greater computational expense.

The agent needs to be able to perform planning that incorporates models of its internal states and external environment. These are deliberative processes and are associated with explicit motivations.

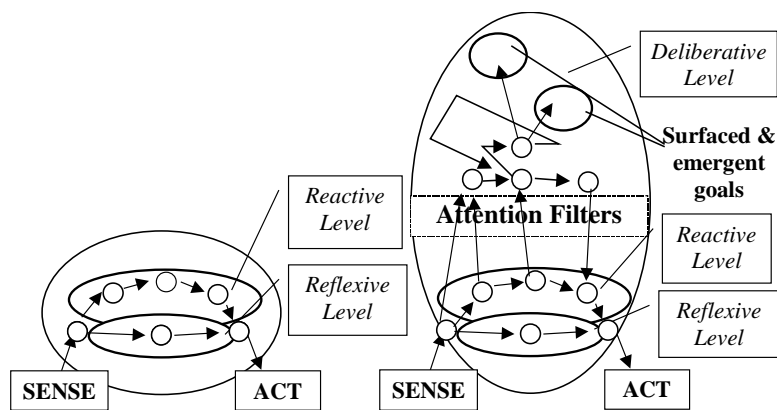
## **MOTIVATION IN COMPUTATIONAL AGENTS**

Motivations are a disposition to assess situations in certain ways and involve perception of problematic events and states, representations and paths to modified states of affairs. This research (Beaudoin & Sloman, 1993; Davis, 2001b) has identified a number of subtypes.

- Goals can be of several sub-types. Quantitative goals are those talked about in control theory, and tend to involve negative feedback. Qualitative goals are the type most used in agents and involve relations, predicates, states, behaviours etc. Hybrid goals are some mix of these two;
- Attitudes are predispositions to respond or act (either internally or externally) to specific (perceptual or cognitive) cues and can involve intricate collections of beliefs, motivators and other control states, for example the selfless and selfish traits of the agents in (Davis, 1997). Similar control states are linked to persona in other work (Sloman, 2001).
- Desires relate objects, agents or events in an agent's world to plausible states of that world. Impulses are transient desires which when acted on give rise to non-transient changes internal to the agent and/or in the external world.

Psychological definitions of emotion can be found (Wilson & Keil, 1999) that refer to both cognitive (appraisal) and physiological factors (reinforcers and valencing). The processes leading to the experience of emotions (in humans) are neither bottom-up nor top-down – they are both

and more. Emotions are experienced as a result of the interactions within and with a synergistic information processing architecture (Davis, 2001a). In short emotions are in part mental (appraisal) states and supporting and causal (valencing) processes. Any computational model of emotion must attempt to meet similar specifications. The stripped down architecture presented here makes use of causal (valencing) processes (i.e. drives) and appraisal processes (i.e. motivations). We can therefore use experimentation with such a process architecture to answer the question of whether such a minimal architecture is capable of supporting recognizable emotive qualities. The final section of this chapter addresses this question.



*Figure 5. Two Action Selection Architectures . Agent0/Agent1 (left) with drives and Agent2 (right) with explicit motivational structure processing. Agent2 is an example of the tightly coupled processes shown in the centre of Figure 1. Both modelled in figure 4.*

The architecture shown on the right hand side of figure 5 adds explicit motivation processing at the deliberative level to the drive model of Agent1. The signal strength associated with a drive means that calls to build explicit motivational constructs for specific drives can have a quantitative (signal) and qualitative (type) component. Motivational constructs can be likened to hybrid goals. The full motivator structure is shown in table 1. This is implemented using the

object-oriented paradigm with control and support processes modeled using methods and rulesets in the SIMAGENT toolkit (rulesets RS6 and RS7 in figure 4).

*Table 1. The components associated with motivator structures.*

Component	Meaning
Actors and Entities	Other agents (actors) and objects referenced by this motivator
Belief Indicator	Indication of current belief about the status of semantic content P: e.g. true, partially true, false.
Commitment Status	The current status of the motivator, e.g. adopted, rejected, undecided, interrupted, stalled, unconsidered, completed.
Decay Function	Defines how insistence decreases while motivator is not adopted.
Dynamic State	The process state of the motivator e.g. being considered, nearing completion etc.
Emotional Key	Processing keys to the emotions and their situational triggers for the motivator.
Importance Value	Importance (e.g. neutral, low, medium, high, unknown). This may be intrinsic or based on an assessment of the consequences of doing or not doing
Insistence Value	Heuristic value determining interrupt capabilities. This should correspond to a combination of the motivator's importance and urgency.
Intensity	This influences the likelihood of (continuing) to being acted on.
Management Information	The state of relevant management and meta-management processes.
Motivational Attitude	The motivator's attitude to semantic content P : make true, keep true, make false etc.
Plan Set	Possible plan or set of plans for achieving the motivator.
Rationale	If the motivator arose from explicit reasoning – motivators need not.
Semantic Content	A proposition P denoting a possible state of affairs, which may be true or false
Urgency Descriptor	How urgent is this descriptor – this may be qualitative (e.g. high, low) or quantitative (for example a time-cost function).

Not all the components of this computational object need to be instantiated for any specific agent class or drive type. Many of the components act as keys that influence processing elsewhere in the agent architecture of figure 1. For example the Emotion Key component can take single word values (e.g. Fear, Hate). This does not imply that emotion is a single linguistic term. Rather these linguistic keys prime and influence related processing. Therefore in this architecture emotion is a

control state distributed across a number of constructs and construct handling processes and any supporting processes.

Consider the maundering scenario described above. The reactive agent will either flee from the predator (Signal Strength Model) or move towards the mean position of the three entities (Behaviour Sum Model). The deliberative agent has other possibilities. Motivational structures can be instantiated that correspond to the following three drives based on figure 3:

- Flee From Predator at location(X,Y) with signal strength 0.75
- Move to Energy Source One at location(Xf1,Yf1) with signal strength 0.5
- Move to Energy Source Two at location(Xf2,Yf2) with signal strength 0.5

Table 2. The instantiated motivator structure for Flee Predator.

<b>Component</b>	<b>Value (initial)</b>	<b>Value (appraisal 1)</b>	<b>Value (appraisal 2)</b>
<i>Actors</i>	Predator1	Predator1	Predator1, Food1
<i>Belief Indicator</i>	True	True	True
<i>Status</i>	Unconsidered	Considered	Adopted, Merged
<i>Decay Function</i>	Null	Null	Null
<i>Dynamic State</i>	Null	Postponed	Active
<i>Emotional Key</i>	Fear	Fear	Anxiety
<i>Importance Value</i>	Null	High	High
<i>Insistence Value</i>	0.75	0.75	0.75
<i>Intensity</i>	Null	Medium	Medium
<i>Management</i>	Null	Null	Combined(M1,M2)
<i>Attitude</i>	Make true	Make true	Make true
<i>Plan Set</i>	Null	Null	Plan(set10, bypass)
<i>Rationale</i>	Drive:Fear	Drive:Fear	Drive:Fear, Drive:Hunger
<i>Content</i>	Flee(Predator1,X,Y)	Flee(Predator1,X,Y)	Flee(Predator1,Xp,Yp) GoTo(Food1,Xf1,Yf1)
<i>Urgency</i>	Null	Null	High

Table 2 gives examples of this structure as created for the Flee From Predator drive in the current situation with the initial values given in column 2. Similar structures are created for the other two (Move to Energy Source) drives. The deliberative processes associated with motivation management are initiated by an alarm call from the posting of these structures to the agent's

motivator database. In this simple architecture all these motivators are given an initial appraisal, leading to the changes in value as shown in column 3 of table 2. The motivator database is then ordered on the basis of the importance value for each motivator. Importance is given a fuzzy-valued descriptor (low, medium or high) based on an appraisal of insistence, emotion key and content. In the simplest motivation architectures the top-most item from the database would be selected and plan set 5 (flee) adopted as behaviours to be activated. A more sophisticated approach can look to see how motivators can be combined, in a manner similar to goal conflation in teleo-reactive planning (Nilsson, 1994). A planning module is used to find plans that combine plan-sets associated with pair-wise combinations of motivators. The ordering on the motivator database dictates the combination order. Motivators that can be combined are subsumed into the highest motivator with the second deleted; in an agent with a more extensive motivator management architecture, the creation of new motivators does not delete existing ones. Any change to the motivator database results in further motivator appraisal. The fourth column in table 2 shows the result of combining the Flee Predator with GoTo Energy-Source1 motivators with a suitable new plan-set. This motivator is adopted, made active and its plan set pursued as a series of movements in the environment that allow the agent to stay away from the predator but still move to and consume one of the energy sources.

Such architectures for (single) agents have been subject to research for a number of years (Beaudoin & Sloman, 1993; Davis, 1997). A distributed version is now being re-developed based on the sketches in figures 1 and 5 (Davis, 2001b; Nunes, 2001). This builds upon the idea of loose and tight coupling of agents as described in earlier work. Rather than build many agents each with their own motivation appraisal and management processes, agents in the same team

(Nunes used five-aside-football as the application domain) initiate motivational structures which are passed to an abstract agent responsible for coordinating the motivations of alike agents. In effect agents in the same herd (or team) share the computationally expensive deliberative processing associated with motivator appraisal. There are a number of reasons why such a direction may prove beneficial. Agent communication overheads associated with motivation coordination are computationally less expensive than multiple agents each with their own motivation management capabilities. Furthermore, given that we require something other than emergent cooperation these latter agents still need to communicate with each other about their adopted motivations.

## **DISCUSSION**

Extending the capabilities of implementations based on the earlier designs was problematic without re-engineering the entire implementation – a process which could take years. With the distributed model, extra agencies could be introduced with these further capabilities. Furthermore difficulty was experienced in introducing learning and adaptation into the design and implementation. It seemed that control mechanisms spanning the entire agent architecture would need changing. Other research has identified the same problem (Franklin, 1997). With the distributed model specific agents can change as the overall architecture adapts. Learning associated with specific capabilities or knowledge is focussed on those agents responsible for those capabilities or knowledge. An analogy can be drawn between this model and with changes observed to classifier rules in Holland's architecture (Holland, 1975). The distributed agent requests those elements of itself (i.e. its component agents) that are associated with the current learning task to modify themselves. Changes to the global agent are in effect mapped onto

changes to agents with a specific process loci. There are many challenges to be faced with this direction.

Currently we are looking at the nature of communication between agents with shared motivations, and are using distributed blackboards (Nunes, 2001) as a metaphor to the global workspace theory that Baars (1988) describes. This approach is being used by others for example IDA (Franklin, 2000). Figure 1 provides the perspective for the experiments with alternative designs for agents with the mechanisms necessary for us to address the stated research objectives. Although (ultimately) many different agent types will be required, no single agent type will be much more complicated than that described for the right hand side of figure 5. Deliberative agents are awoken by alarms when required. In one such architecture agents develop motivator structures and pass them to a shared motivator appraiser agent. This agent has sufficient knowledge of the communicating agents that it can select motivators as appropriate. Other agents can be called on to perform specific tasks by the reactive agents or by other deliberative agents as required. The motivator appraisal described above for the maundering reactive agent could therefore be applied to many agents making use of the motivator-initiating reactive agent, one motivator appraisal agent and a planning agent.

The design philosophy described above is proving useful in developing agent applications. We are finding that developing design methodologies that can cope with the sometimes elusive concepts and ideas associated with cognitive science is of benefit to intelligent agent modeling in more constrained domains such as enterprise management (Davis, 2000), clinical diagnosis and e-commerce (Luo et al, 2002). The applications implemented so far are making use of what we have learnt from developing adaptive agents for cognitive science research.

The cognitively oriented software agents described in this paper are helping us understand the concepts underlying drives and motivation. We firmly believe that no matter how eloquent philosophical or psychological theories and models, the plausibility of the information processing and other mechanisms embodied in these theories can only be validated through the development of computational models. The type of experiments described here and elsewhere (Davis, 1997, 2001a, 2001b; Nunes, 2001) exemplify this approach. We have designed and implemented agents that display motivational qualities and address important questions about the nature of emotion and autonomy. Meta-agent architectures and adaptive agents provide the design and implementation tools necessary to pursue such lines of inquiry. The work in this chapter highlights the relation between agent architectures and the nature of drives and motivations. It shows that we can differentiate between drives and other forms of motivational control state at a theoretical, design and computational level. With the advent of affective computation (Picard, 1997) and the growing sophistication of software systems, the types of control structures discussed in this chapter may well be required in future agent applications. We have shown that agents demonstrate sometimes unwanted qualities, i.e. control dithering and maundering. We have shown that by improving the processing structures associated with drives and goals (i.e. the use of motivations), deliberative control of such emergent and affective states can be achieved. Other experimentation has shown that oscillations in this deliberative control of emergent affective states can occur. The inclusion of the reflective (meta-agent) layer in the agent architecture provides the means to control this phenomena. Earlier it was been argued that the simple Agent0/Agent1/Agent2 architecture is capable of demonstrating affective qualities such as maundering. Such results are in agreement with other researchers (Scheutz & Sloman, 2001). Now consider whether we have demonstrated emotive qualities in our research agents. Oatley

and Jenkins (1996) define emotion as “*a state usually caused by an event of importance to the subject. It typically includes (a) a conscious mental state with a recognizable quality of feeling and directed towards some object, (b) a bodily perturbation of some kind, (c) recognizable expressions of the face, tone of voice, and gesture (d) a readiness for certain kinds of action*”.

Others, for example (Frijda, 1986), give similar definitions. This school of thought permits basic emotions, of which Fear is one. Fear is defined as the physical or social threat to self, or a valued role or goal. Do any of the above agents satisfy these descriptions? Prey agents recognize the localized presence of a predator as a threat. Agents with motivations produce motivational states (structures and processes) to deal with this identified threat (part a of the definition). Even the Agent0 base plinth displays a perturbation as a movement away from the threat (part b of the definition), which is also an indication of readiness to respond to the threat. The agents with motivations also have plans to avoid the threat (part d of the definition). It therefore follows that these agents are in fact capable of displaying a limited repertoire of emotive qualities, even if in a rather shallow manner.

In pursuing this line of research we have found it necessary to revisit foundational principles in agent theory such as for instance autonomy, flexibility and adaptability. Domain models of autonomy sometimes raise problems in the mapping from theory to design to implementation. A domain model of autonomy in economics, social theory, cognitive science or artificial life may be at odds with or compromised by the notion of autonomy underpinning the agents in a computational toolkit. Such compromises can lead to the design of agent systems with inherent conflicts. This is an issue that needs to be addressed by research into agent software engineering. By designing agents with the qualities described in this chapter an agent is given the means to

represent and reason about these conflicts when they do arise. This research continues to raises questions about what agent is, what a mind is, and what are emotion and motivation.

## **REFERENCES**

- Agre, P. & Chapman, D. (1987). PENGI: An implementation of a theory of activity. *Proceedings of AAAI-87*, 268-272. Seattle. WA.
- Baars, B.J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Bates, J., Loyall, A.B. & Reilly, W.S. (1991). Broad agents. *SIGART BULLETIN*. 2(4).
- Beaudoin, L.P. & Sloman, A. (1993). A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge & A. Ramsay (Eds.), *Prospects for Artificial Intelligence*, IOS Press.
- Boden, M.A. (1996). *The Philosophy of Artificial Life*. Oxford University Press.
- Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence*. 47, 139-159.
- Castelfranchi, C. (1995). Guarantees for autonomy in cognitive agent architectures. In M. Wooldridge & N.R. Jennings (Eds). *Intelligent Agents*. Springer-Verlag. 56-70.
- Chapman, C.R. (1996). Limbic processes and the affective dimension of pain. In: G.Carli & M. Zimmerman (Eds.). *Towards the Neurobiology of Chronic Pain*. 110, 63-81.
- Ciancarini, P. & Wooldridge, M. (2001). *Agent-Oriented Software Engineering*, Springer-Verlag.
- Davis, D.N. (1997). Reactive and motivational agents. In J. P. Muller, M. J. Wooldridge & N. R. Jennings (Eds.), *Intelligent Agents III*. Springer-Verlag.
- Davis, D.N. (1998). Synthetic Agents: Synthetic Minds? *Frontiers in Cognitive Agents. IEEE Symposium on Systems, Man and Cybernetics*. San Diego.
- Davis, D.N. (2000). Agent-Based Decision Support Framework for Water Supply Infrastructure Rehabilitation and Development. *International Journal of Computers, Environment and Urban Systems*. 24, 1-18.
- Davis, D.N. (2001). Multiple Level Representations of Emotion in Computational Agents. *AISB'01 Symposium on Emotion, Cognition and Affective Computing*. University of York.

- Davis, D.N. (2001). Control States and Complete Agent Architectures. *Computational Intelligence*. 17(4).
- Epstein, J.M. & Axtell, R. (1996). *Growing Artificial Societies*. MIT Press.
- Erman, L.D. & Lesser, V.R. (1975). A multi-level organisation for problem-solving using many diverse cooperating sources of knowledge. *Fourth International Joint Conference on Artificial Intelligence (IJCAI-75)*, 483-490.
- Ferber, J. (1999). *Multi-Agent Systems*. Addison-Wesley.
- Frankel, C.B. & Ray, R.D. (2001). Competence, Emotion and Self-Regulatory Architecture. *AISB'0 Symposium on Emotion, Cognition and Affective Computing*, University of York.
- Franklin S.P. (1997). Autonomous Agents as Embodied AI. *Cybernetics and Systems*. 28(6), 499-520.
- Franklin, S.P. (2000). A “consciousness” based architecture for a functioning mind, *AISB'00 Symposium on How to Design a Functioning Mind*, University of Birmingham.
- Frijda, N. (1986). *The Emotions*. Cambridge University Press.
- Gibson, J.J. (1986). *The Ecological Approach to Visual Perception*, LEA Press.
- Hanks, S., Pollack, M.E. & Cohen, P.R. (1993). Benchmarks, Test-beds, Controlled Experimentation, and the Design of Agent Architectures. *AI Magazine*. 14(4), 17-42.
- Hayes-Roth, B. (1993). Intelligent control. *Artificial Intelligence*. 59, 213—220.
- Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.
- Husbands P., Harvey, I. & Cliff, D. (1993). An evolutionary approach to situated AI. In A.Sloman, D.Hogg, G. Humphreys, D. Partridge & A. Ramsay (Eds.), *Prospects for Artificial Intelligence* (pp 61-70). IOS Press.
- Luo, Y., Davis, D.N. & Liu, K. (2002). Combining KADS with Zeus to Develop a Multi-Agent E-Commerce Application. *International Journal of Electronic Commerce Research* (In Press)
- Nilsson, N.J. (1994). Teleo-reactive programs for agent control. *Journal of Artificial Intelligence Research*. 1, 139-158.
- Nunes, H.A. (2001). *Investigation of motivation in agents using the simulation of 5-a-side football*. M.Sc. Research Thesis, Computer Science, University of Hull.

- Nwana, H., Ndumu, D., Lee, L. & Collis, J. (1999). "ZEUS: A toolkit for building distributed multi-agent systems", *Applied Artificial Intelligence Journal* 13(1), 129-186.
- Oatley, K. & Jenkins, J.M. (1996). *Understanding Emotions*. Blackwell.
- Ortony, A., Clore, G.L. & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.
- Picard, R. (1997). *Affective Computing*, MIT Press.
- Ryle, G. (1949). *The Concept of Mind*. University of Chicago Press. (Reprinted 1984).
- Scheutz, M. & Sloman, A. (2001). Affect and Agent Control: Experiments with Simple Affective States. In N. Zhong, J. Liu, S. Ohsuga & J. Bradshaw (Eds.), *Intelligent Agent Technology: Research and Development* (pp200-209). New Jersey: World Scientific.
- Schreiber, G., Weilinga, B. & Breuker, J. (1993). *KADS: A Principled Approach to Knowledge-based Systems*. Academic Press, London.
- Simon, H.A. (1967). Motivational and emotional controls of cognition. In: H.A. Simon, *Models of Thought*. Yale University Press.
- Sloman, A. & Croucher, M. (1987). Why Robots Will Have Emotions. *IJCAI7*, Japan, 197-202.
- Sloman, A. (1993). The mind as a control system, In C. Hookway & D. Peterson (Eds.), *Philosophy and the Cognitive Sciences*. Cambridge University Press.
- Sloman, A. & Logan, B. (1998). Cognition and Affect: Architectures and Tools. *Proceedings of the Second International Conference on Autonomous Agents (Agents '98)*, ACM Press.
- Sloman, A. (2001). Varieties of Affect and the CogAff Architecture Schema. *AISB'01 Symposium on Emotion, Cognition and Affective Computing*. University of York.
- VanLehn, K. (1991). *Architectures For Intelligence*. LEA.
- Wheeler, M. (1997). Cognition's coming home: the reunion of life and mind. *Fourth European Conference on Artificial Life*. MIT Press. 10-19.
- Wilson, R.A. & Keil, F.C. (1999). *The MIT Encyclopedia of the Cognitive Sciences*, MIT Press.
- Wooldridge, M. & Jennings, N.R. (1995). *Intelligent Agents*. Springer-Verlag. 56-70.