Minds have personalities - Emotion is the core

Darryl N. Davis

Neural, Emergent and Agent Technology Research Group,

Department of Computer Science, University of Hull,

Kingston-upon-Hull, HU6 7RX, U.K.

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

Abstract

There are many models of mind, and many different exemplars of agent architectures. Some models of mind map onto computational designs and some agent architectures are capable of supporting different models of mind. Many agent architectures are competency-based designs related to tasks in specific domains. The more general frameworks map across tasks and domains. These types of agent architectures are capable of many cognitive competencies associated with a functioning mind. However, there is a problem with many of these approaches when they are applied to the design of a mind analogous in type to the human mind – there is no core other than an information processing architecture. As any specific architecture is applied to different domains, the information processing content (knowledge and behaviours) of the architecture changes wholesale. From the perspective of developing intelligent computational systems this is more than acceptable. From the perspective of developing or simulating functioning (human-like) minds this is problematic – these models are in effect autistic. This paper presents an emotion-based core for mind. This work draws on evidence from neuroscience, philosophy and psychology. As an agent monitors its internal interactions and relates these to tasks in its external environment, the impetus for change within itself (i.e. a need to learn) is manifested as an unwanted combination of emotions (a disequilibrium). The internal landscape of emotion, control states and dispositions provides a basis for a computational model of personality (and consciousness).

1. Introduction

For much of its history, cognitive science has positioned emotion as the poor relation to cognition. For many emotion is the Achilles' heel of reason. This paper takes a stance on (human-like) minds that places emotion as the core. From a computational perspective, the impetus for this research is the inadequacy of earlier work on the modelling of motivation (Davis 1996) to adequately contain aspects of cognitive functioning. This paper takes a trajectory through work from neuroscience on what parts of the central nervous system play a role in emotions, research from psychology and analyses from philosophy. This paper will not give a definitive definition of emotion but look to argument and finding agreement from a number of sources in ascribing the role of emotion in functioning minds. A sketch of a computational theory of mind (primarily from the agent perspective) will be then be considered in the light of this evidence. This leads onto the presentation of preliminary experimental work that models emotions as the core of a computational architecture of a mind.

2. Emotions and the mind

The nature of emotions and the relation to thought have been analysed since the dawn of western civilisation. Plato degrades them as distorting rationality. Aristotle denotes long tracts to their categorisation and impact on social life. For Darwin emotions in adult humans are a by-product of evolutionary history and personal development.

Here the definition of emotions as "*…examples of non-problem-solving non-behaviour*" (Gunderson 1985:72) is completely rejected. Merleau-Ponty supposes humans are moved to action by disequilibria between the self and the world. Emotion plays a large role in initiating and providing descriptors for such disequilibria. Emotion is a primary source of motivation. Criminal law recognises the importance of emotions in

differentiating between voluntary manslaughter (occurring in the heat of passion) and murder (involving malice aforethought and deliberate suspension of control). French law takes this further with its concept of crimes of passion. However to consider emotions solely as an emergent quality of mental life that undermines reason and rationality is "*a vehicle of irresponsibility, a way of absolving oneself from those fits of sensitivity and foolishness that constitute the most important aspects of our lives*" (Solomon 1993:131-132). Emotions are "*a subjective strategy for the maximisation of personal dignity and self-esteem*" (Solomon 1993:222). Schenck (2000) in his study of the role of music suggests that there are resource and motivation problems associated with this tension between emotions and cognition and that "*we are rational only when we have the time, or the inclination to be so*". Much of psychopathology and psychiatry is concerned with understanding how minds dysfunction. Depression, mania and phobias are often associated with affective disorders. Much of the treatment of depression revolves around identifying and correcting the sources for the emotions of fear and anxiety. Damasio's text (1994) details how physiological damage to the prefrontal cortex, the limbic system (in particular the amygdala) and the afferent pathways that connect the two areas result in emotional dysfunction, personality change and a loss of reason (dissociation). Again emotions play an important role in the executive aspects of cognition, i.e. judgement, planning and social conduct. Goleman (1995) terms this emotional intelligence - it appears to be very similar to what others (see Spaulding 1994) term social intelligence. Emotion has many functions including the valencing of thoughts related to emerging problems, tasks and challenges in terms of emotional intensity and emotion type, as in for example directing attention to aspects of internal and external environments. Such a function is a precursor to problem solving. Many researchers have written on the importance of emotion for motivation (Simon 1979; Spaulding 1994), memory (Rolls 1999), reason (Damasio 1994) and learning. Solomon suggests that "*there is no ultimate distinction between reason and passion*", and that together the two provide more than an understanding of experience, they constitute it. In short emotion has a central role in a functioning mind.

The conjecture cognitive scientists need to face is whether the computational modelling of human-like minds is possible without a silicon/digital analogue to human-like emotions. Research into producing computational cognition may lead to the development of intelligent problem-solvers of many types (e.g. ACT, AIS, SOAR), but the simulation of the human mind requires other categories of intellectual processes. Much of cognitive science and artificial intelligence adopts a modular approach to cognition. If vision, memory, attention, language can be solved, an artificial brain can be built. Such an artefact will perceive, reason and act in its world, relating current to past events, focusing on cognitive salient events in that world. It will interact with and represent parts of its external environment but it will have no internal environment and no sense of self. Without emotions it will be diagnosed as autistic! This approach to cognitive science is one that Harré (1994) argues against - the individual as passive observer of the computational processing that is that person's cognition. Cognition is part of the mental repertoire – perhaps a large part but it is not the entirety of the mind. The efficacy of its use depends on the mind it serves. In looking for general principles to the functioning of mind, cognitive science has perhaps neglected those aspects of mental life that give rise to individual differences. This is perhaps understandable as science looks to general principles. However a redress is called for, and to understand how a mind functions, general principles that also explain individual differences need to be found.

An alternative stance is to place emotion at the core of mind. This core gives rise to episodic states (e.g. feelings), trajectory states (e.g. moods and dispositions) and (semi-permanent) endogenous states (e.g. personality). Personality traits lasting years (or a lifetime) are usually tightly bound to qualities of emotions. To rephrase a previous revolution in artificial intelligence: *human-like intelligence requires embodiment of the supporting computational infrastructure not only in terms of an external environment but also in terms of an internal (emotional) environment.*

3. Psychology and emotion

Over the last hundred years of psychology (from James onwards) the study of emotion has waxed and waned with theories of emotion typically rooted in discussions of physiological and non-rational impulses and drives. An exception is the "cognitive" school of emotion dating from Paulhan (1887) through to Schacter and Singer's (1962) influential experiments with adrenaline and the effect of social context on emotive appraisal. A standard introduction to psychology from the 1970s (Lindsay and Norman 1972) summarises much of the experimental work on emotions in suggesting that emotional states are manipulable through cognitive processes (in particular expectations), physiological states and environmental factors. They conclude that cognition (particularly memory, motivation, attention and learning) and emotions are intimately related. In Newell's seminal work on cognition (Newell 1990), emotion is not indexed and is only discussed in any length in relation to social aspects of a cognitive agent in the final chapter. Although Newell acknowledges this, it reflects a trend in cognitive science to place

emotion as subordinate to rationality and cognition. Despite pointers to the importance of understanding emotion for cognitive science (e.g. Norman 1985), cognitive science all too readily follows as a modern day Stoic successor to Plato in minimising the role of emotion. A leading volume on the dynamics approach to cognition (Port and Van Gelder, 1995) is no exception – particularly odd if emotion is viewed as the *flow and change* of cognitive predisposition over time and across occasion (Lazurus 1991).

Ortony et al (1988) consider cognition to be the source of emotion, but that unlike many other cognitive processes, emotions are accompanied by visceral and expressive manifestations. They consider valence (i.e. positive-neutral-negative) and appraisal (cognitive reflection of these valences) as the primary basis for describing an emotion. They differentiate emotions from non-emotions on the basis of whether a valenced reaction is necessary for that state. However, non-emotion states (e.g. abandonment) can give rise to causal chains of emotive reactions leading to highly valenced (emotive) states. They suggest that there are basic classes of emotion related to valenced states focussed on events (pleased vs. displeased), agents (approving vs. disapproving) and objects (liking vs. disliking). Specific emotions are instances and blends of these types and subclasses. Emotions of the same type have eliciting conditions that are structurally related. They reject the idea of emotions such as anger and fear being fundamental or basic emotions. The cognitive processing that appraises emotions is goal-based and resembles the type of processing and structures discussed in motivation for autonomous agents (e.g. Beaudoin and Sloman 1993, Davis 1996).

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) recognisable expressions of the face, tone of voice, and gesture (d) a readiness for certain kinds of action*". Others (e.g. Frijda 1986) give similar definitions. A number of other psychologists (e.g. Power and Dalgleish 1997) appear to be in agreement in defining what are basic emotions:

- Fear defined as the physical or social threat to self, or a valued role or goal.
- Anger defined as the blocking or frustrations of a role or goal through the perceived actions of another agent.
- Disgust defined as the elimination or distancing from person, object, or idea repulsive to self and to valued roles and goals.
- Sadness defined as the loss or failure (actual or possible) of a valued role or goal.
- Happiness defined as the successful move towards or completion of a valued role or goal.

They suggest that these five suffice as the basic emotions as they are physiologically, expressively and semantically distinct. There are cases for other emotions to be considered as further basic emotions. From a perspective of classifying emotions using distinctive universal signals, i.e. expressions (Ekman & Davidson 1994), surprise is included in this fundamental set. However, from the perspective of classifying emotions based on distinctive physiological signs (see Power and Dalgleish 1997), the basic set is reduced to fear, anger, disgust and sadness.

Rolls (1999) presents a different perspective on the psychology of the emotions. Brains are designed around reward and punishment (reinforcer) evaluation systems. While this can be seen as analogous to the valenced arousal states in the Ortony et al. theory, the reinforcers are precursors to any specific emotion. Rather than reinforcing particular behavioural patterns of responses (behaviourism), the reinforcement mechanisms work in terms of cognitive activity such as goals and motivation. Emotions are states elicited by reinforcers. These states are positive when concerns (goals) are advanced and negative when impeded. Again, there is an overlap with the perspectives of Power and Dalgleish, and Oatley and Jenkins. These states are more encompassing than those states associated with the mere feelings of emotion. This aspect is considered further in Wollheim's analysis of the emotions. Emotions have many functions (Rolls lists ten) including the priming of reflexive behaviors associated with the autonomic and endocrine system, the establishment of motivational states, the facilitation of memory processing (storage and control) and maintenance of the "*persistent and continuing motivation and direction of behavior*". In effect Rolls suggests that the neuropsychological evidence supports the conjecture that emotions provide the glue that binds the multitude functions of mind.

4. Philosophy and emotion

Wollheim (1999) distinguishes two aspects of mental life in his analysis of emotion: the phenomena of mental states and mental dispositions. Mental states are temporally local to their initiating event and transient, being relatively short-lived - sometimes instantaneous. Mental states can reoccur frequently to give the impression of a continuous state. Mental dispositions can more long-lived (sometimes over a lifetime) – they are temporally global - they have histories. Mental states and dispositions are causally related. Mental states can instantiate and terminate mental dispositions.

Mental states can reinforce and attenuate mental dispositions. Mental dispositions can also facilitate mental states. Both mental states and dispositions have a psychological reality. Impulses, perceptions, imaginings and drives are mental states. Beliefs, knowledge, memories, abilities, phobias and obsessions are examples of mental dispositions. Three very general properties characterise these two types of mental phenomena: intentionality, subjectivity and three exclusive grades of consciousness (conscious, preconscious and unconscious). Both mental states and dispositions have an intentional quality – i.e. they are related or directed to either internal or external events. Wollheim suggests that subjectivity be only associated with mental states – mental dispositions can only be indirectly experienced through the mental states in which they are manifest. It is in highlighting the very differences between mental states and dispositions that Wollheim makes use of the emotions. Emotional states differ from emotional dispositions. Emotions are preconscious mental dispositions and cannot be directly experienced. What can be experienced are feelings or perceptions of emotion (mental states) associated with mental dispositions. While the two can be causally interrelated this need not be the case. Mental dispositions are preconscious (and in some cases unconscious) traits. We can become aware of (aspects of) them though training (e.g. yoga) or therapy and in doing so make parts of the preconscious mind conscious. In everyday functioning however the conscious mind is aware of mental states and relates these to personal histories and intended futures – the current, past and intended states of being.

From an computational perspective on the philosophy of mind, Sloman has for many years considered that intelligent machines will necessarily experience emotion (-like) states (Sloman and Croucher 1987). Following on from the work of Simon (1979), his developing theory of mind and the nature of problem solving considers how in attempting to achieve multiple goals (or motivators) perturbant (emotion-like) states ensue (Wright et al 1996). These perturbant states will arise in any information processing infrastructure where there are insufficient resources to satisfy current and prospective goals. Sloman (1987) tends to describe emotion in terms of disturbances of mental processes (the Achilles heel again!). Like Wollheim, Sloman differentiates between episodic and persistent mental phenomena, both of which can carry emotional constituents. More recently his architectures for functioning minds include primary, secondary and tertiary emotions (Sloman 1999). Primary emotions are analogous to arousal processes in the theories introduced above (i.e. they have a reactive basis). Secondary emotions are those initiated by appraisal mechanisms (i.e. they have a deliberative basis). Tertiary emotions are cognitive perturbances - negatively valenced emergent states - arising from (typically goal or motivator) conflicts in an information processing architecture. Tertiary emotions arise from the interaction of emotions and other cognitive processes (e.g. motivation) at the deliberative layer. In many situations these perturbant states arise through resource inadequacy or mismanagement while pursuing multiple and not necessarily incompatible goals. While the work that follows certainly builds upon some of these ideas, this framework seems flawed. Perhaps the differentiation that Sloman makes between these emotions can be more easily explained in terms of the different categories of processing that the mind performs over its different layers. A secondary emotion is an analogous state (or disposition) to a primary emotion but seemingly perceived in a different manner due to the characteristics of the processing at the different layers. In visual perception terms, the red object that swept past our visual senses, causing a startled (reactive) response, that disturbs ongoing thought and behaviour patterns, is the same red object that is subsequently perceived as a rose petal blown by the wind from a nearby shrub in the garden.

5. Theoretical Framework

The theoretical framework presented here builds on those aspects of agreement in the work presented above. It revisits an earlier (computational) architecture of mind and emphasises the interplay of cognition and emotion through appraisal, motivation and niche space. Psychological definitions of emotion have been presented that refer to cognitive (appraisal) and physiological factors (arousal), and the valencing of emotive states and reinforcers as precursors to emotional arousal. 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 that includes (at least) the endocrine system, the limbic system, and the cortices. Emotions, in socio-biological agents, are in part mental (appraisal) states and supporting (valencing) and causal (reinforcer) processes. Any computational model of emotion must attempt to meet similar specifications, and address the differentiation in mental phenomena that Wollheim makes. In moving towards a model of emotion that will be computationally tractable, the extent of the model will be initially (at least) minimised. A minimal model of emotion enables the model to be used as the core to an agent-based model of the mind.

Earlier research on agents focussed on an architecture that supports motivation (Davis 1996). The architecture (sketched in figure 1) emphasises four distinct processing layers: a reflexive layer that is analogous to

the autonomic systems in biological agents, a reactive (preconscious) layer, a deliberative layer and a reflective layer. This broad picture of the mind has high level and low level processes co-existing and interacting in a holistic manner. Hence motivator processing, planning, decision-making and other cognitive processes are not merely abstract but exist in relation to other automatic, autonomous and adaptive processes. The entirety of the agent's processing exists in relation to the agent's environmental stance; i.e. what objects, agents and events are occurring in the environment and how they affect the goals and motivations of the agent. The two lower layers relate to *pre-attentive* processes and are capable of supporting innate and learnt environmental competencies and (internal and external) behaviours. Perception of and action upon the external environment is mediated primarily through these two layers. The third (deliberative) layer relates to the types of things discussed in most cognitive science, for example (Newell 1990). This does not preclude a non-symbolic implementation of this layer. The fourth layer, the reflective qualities, serves to monitor the overall behaviour of the agent. In particular, the role of the reflective layer is to identify and act on out-of-control behaviours, whether internal, external, deliberative or reactive. This (reactive, non-deliberative) meta-management level processing is considered to be the most abstract level of processing. If it were not, there is a requirement for the reflective processes to be monitored in turn - this in effect would lead to an infinite regress.

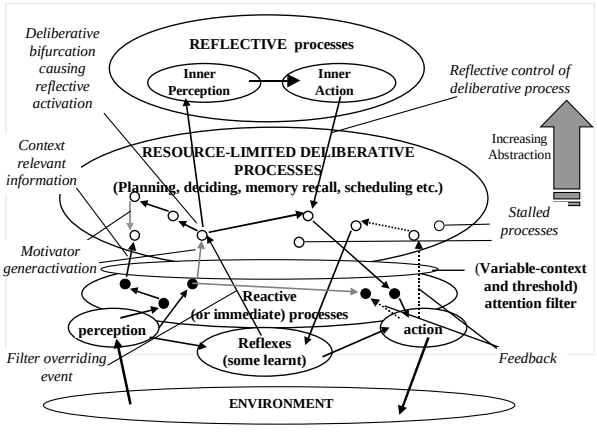

Figure 1. Sketch of an architecture of a mind.

Control suggestions from reflective layer do not always override processes originating and ongoing in the other layers. The behaviour of an intelligent cognitive agent is not controlled by any of these layers in isolation. Behaviours at the reactive level may preclude processes at or actions motivated by the deliberative or reflective layers. Processes over any specific combination of layers may arise as a result of an agent attempting to manage control states originating in any of the layers. Where decision processes related to possibly antagonistic behaviours are not cleanly integrated, there is the very real possibility that the agent will experience cognitive perturbance, particularly where the underlying motives are acute (Wright et al 1996). This cognitive perturbance can be described within an emotional context using tertiary emotions (Sloman 1999).

This analysis presents an incomplete picture. In the earlier work the primary analysis of the mind and the resulting computational designs and systems focussed on motivation and goal processing. This analysis was phrased in terms of niche spaces, design spaces and control states. The niche-design space analysis is still valuable tool in designing a functioning mind. However Wollheim's analysis of the mind and emotions, if accepted, will ultimately require a review of the taxonomy used to relate different control states. A deeper analysis of these control states is required, in terms of temporal extent, subjectivity and grades of consciousness. The structures used in modelling motivation incorporated an emotional indicator that corresponds to a deliberative analysis of the motivator and its context. This semantic labelling is insufficient to model emotions. In biological agents emotions are experienced in a conscious, preconscious and physiological sense, and to some lesser or greater extent in terms of post-hoc rationalisation. Over a lifetime, given no cerebral dysfunction, this emotional landscape is navigated in the attempt to achieve life-goals. This can be viewed as moving between neighbouring niche spaces – for example in moving from music student to professional musician. More dramatic changes in desired niche-space are obviously possible. Different trajectories (goal-achieving behaviours) are possible for any such move. Some trajectories while impossible are supported or attended to for any number of reasons. Emotional intensity associated with the preferred niche space (as in the case of grief and the loss of a loved one) is one example. The preferred trajectory between these niche spaces depends on personality and preferred aspects of the emotional landscapes. The emotional landscape is our internal niche space that allows us as biological agents to understand external events, objects and agents in terms of internal (personal) experience. Our biological design (and psychological capabilities and preferences) define the constraints that determine whether any trajectory between niche spaces is possible (or desired).

The emotional landscape that needs to be modelled in building a functioning mind has to address the four layers of the architecture. Figures 2 and 3 present an integrated model of emotion at the core of a simplified version of the architecture given in figure 1. This model is built upon a trajectory through the research presented in the first half of this paper. An agent typically maintains an ongoing (globally temporal) disposition.

The nature of this disposition is (perhaps only temporarily) modified through current goals and motivations. Over time events occur that modify, stall, negate or achieve goals. Such events can occur over all layers of the architecture. These events give rise to reinforcers. The emotion(s) they reinforce depends on their interactions with conscious and preconscious states and dispositions. A valencing component is needed for any emotion. Both the reinforcer and the (preconscious) valences can be modelled using the interval [-1,1] - this interval need not be linear. A discrete version of this interval maps onto the three tokens: negative, neutral and positive. Thirst, hunger, reproduction etc. are physiological and genetic drives, not emotions. These can be associated with reinforcers and be valenced. They can also be associated with motivators – not all motivators need a source in the emotions. The management and success (or otherwise) of these drive-generated motivations can give rise to emotions. There is case for basic emotions. There is considerable agreement that the set of basic emotions includes anger, fear, disgust and sadness. The definitions given above suffice with one exception. Sadness and happiness are antipathetic, being reflections of each other, or extremes on one dimension. Here the term sobriety is used, with sadness and happiness either side of a neutral state. Sobriety is then defined as no change to a valued role or goal. Happiness and sadness are defined as above. A salient feature of the Oatley, Jennings, Power and Dalgleish definitions of emotion is that they are described in terms of goals, roles and expressive behaviours. This enables emotions to be defined over different levels of the architecture using different aspects of motivational behaviours. The type and subclass analysis of Ortony et al. can be used to build upon this basic set of emotions. The resulting four dimensional model is computationally tractable, and maps onto our ideas for the types of cognitive processing (with particular regard to motivation) that occurs in a mind.

Emotional events are temporally short, although emotional states resulting from successive waves of emotional events can be more enduring. Emotions can be casually inter-related and cause other events. Drives and motivations are highly inter-linked with emotions. These can be embodied in some representation (not necessarily semantic) and in effect relate short-term emotive states to temporally global processes. It is suggested that personality traits are focused at the reflective layer and permeate the rest of the architecture, providing the control patterns that stabilise a personality. Personality traits can be seen as dispositions that affect the reflective processes and influence the different categories of cognitive and animated behaviour. Personality becomes an emergent property of the entire architecture and its disposition to favour specific aspects of the possible emotional landscape, and concentrate on tasks that maximise that aspect of the landscape. Personality traits affect and influence the different categories of cognitive and animated behaviour. Moods arise from the interaction of current temporally global niche roles (the favouring of certain aspects of emotion space) and temporally local drives that reflect the current focus of the deliberative processing. Temporally-global drives are those associated with the agent's overall purpose related to its current, possible and desired niche spaces. Temporally-local drives are related to ephemeral states or events within the agent's environment or itself. These can give rise to more enduring motivational states, which may be acted on.

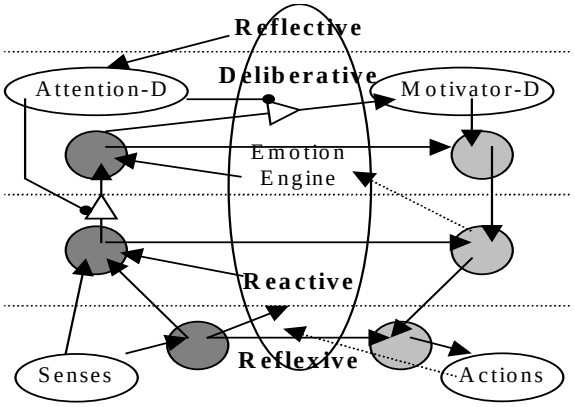

Figure 2. Sketch of the simplified four-layer architecture with emotion as the core. Dark grey circles represent information assimilation and synthesis processes. Light grey circles represent information generation processes that typically mapping into internal and external behaviours. White triangles are filters. Dashed lines represent feedback.

If emergent behaviours (related to emotions) are to be recognised and managed then there must be a design synergy across the different layers of the architecture. Processes at the deliberative level can reason about emergent states elsewhere in the architecture using explicit representations. The reflective processes can classify the processing patterns of the agent in terms of combinations of the four emotions and favoured emotional dispositions. The emotion-changing (reactive) behaviours can be used to pursue a change in emotional disposition. However emotion is not purely top-down processing – as highlighted by Solomon in his differentiation between passion and emotion. Aspects of emotions can be preconscious and, for example, be managed by the autonomic nervous system and its biological substrate (including the endocrine systems). Emotions can move into the conscious mind or be invoked at that level (through cognitive appraisal of agent, object or event related scenarios). Emotions can be instantiated by events both internal and external at a number of levels of abstraction, whether primary

(genetic and/or ecological drives), behavioural or by events that require substantive cognitive processing. In the model in figure 3, intense emotions effectively override the emotion filter causing the forced deliberative consideration of the emotional state. Similar filters are used in the earlier work on motivator generactivation (Davis 1996). The deliberative appraisal of the emotion then acts laterally at the deliberative layer, affecting memory management, attention filters and motivator management. The reflective layer of the mind, which is described entirely in terms of the emotion engine, responds asynchronously to the deliberative phenomena.

6. Experimental computational work

The architecture for a computational mind is based on ideas developed within the Cognition and Affect group at Birmingham (Beaudoin and Sloman 1993; Davis 1996). Rather than reiterate the computational work on the non-emotion aspect of that architecture, here preliminary computational and design experiments with the emotion engine are presented.

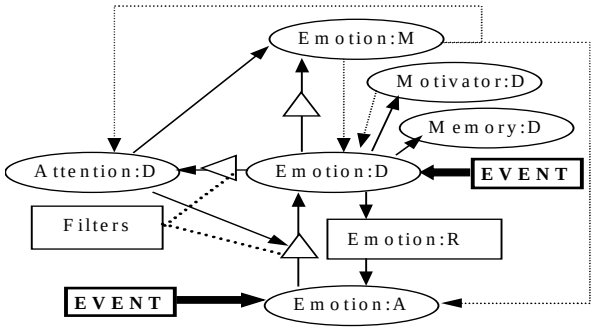

Figure 3. The Emotion Engine for figure 2.

Figure 3 presents a four layer processing model of the emotions. The autonomic processes (Emotion:A) present a base for the model both for dispositional processing and inflection of ongoing dispositions through preconscious events. Such inflections are instantiated by events both external and internal to the agent. The reactive behaviours (Emotion:R) control the functioning of all the preconscious processes. The currently extant Emotion:R behaviours are set by deliberative processes (Emotion:D). The Emotion:M module encompasses the entirety of the meta-management (reflective) processes in this model of the mind. The reflective processes monitor the deliberative appraisal of the Emotion:A processes and the state of the attention filters (managed by Attention:D). The output from Emotion:M provides guidance to the attention management, Emotion:D and the Emotion:A processes. The agent learns to manage its emotions through the development of these five modules. Other aspects of the emotion engine are the placement of deliberative motivator processes, directly affected by Emotion:D. Memory management (Memory:D) is similarly affected.

For a number of reasons the Emotion:A module is modelled using multiple communities of cellular automata. This builds on earlier work (Davis et al 1999) in landscaping decision spaces for the game of Go, and the usefulness of using cellular automata for the modelling of complex social dynamics (Hegselmann and Flache 1998). The behaviours associated with the Emotion:R module govern the internal behaviour of single cells, the communication between adjoining cells in communities and inter-community communication. Different community types have been used. The first experiments (Davis 2000) made use of an insect hive metaphor, with each hive representing an (preconscious) emotional disposition. At the centre of the hive is a (four-dimensional) queen cell that represents the four basic emotions (anger, disgust, fear and sobriety). Each dimension is discretely valenced as positive-neutral-negative. Surrounding the queen cell are four (3-state) drone cells; each mirroring one of the emotions. The remaining (2-state) cells act as filters (guards) or information carriers (worker cells). Further CA communities are being experimented with. The other community type (mobiles) consists of guard and drone cells. This community type represents a reinforcer - a valenced pre-emotive event. Communication between different hives (and input from events outside of the emotion engine at the preconscious level) is by means of the mobile communities. The behaviour of each cell and inter-cell communication is governed by 10 sets of behaviours (50 behaviours in total) plus another behaviour set for inter-community communication. The currently set behaviour from these eleven sets for any hive or hive-mobile combination is selected (as a reactive disposition) by a deliberative (Emotion:D) process. These processes are also responsible for asynchronously monitoring these communities in response to intense hive states and to guidance from the meta-management (Emotion:R) module. Experiments have shown that from any given state, the CA communities rapidly achieved a steady state. By changing the currently extant behaviour set or by communicating with another hive (through the use of a mobile) transitions to the same or other steady states occurs. The CA communities are therefore capable of representing transient and persistent dispositions. The deliberative processes change their emotional disposition (the temporally-global aspect of emotions) and hence the currently extant behaviour set for their hive in response to the reflective processes. The deliberative processes also disturb the motivator and the attention management processes as part of the emotive state appraisal mechanism. Appraisal occurs in response to highly valenced emotive states at the CA communities, feedback from the motivation module or

from events occurring elsewhere in the global architecture at the deliberative level. Memory management (Memory:D) also responds to the Emotion:D processes in order to provide emotional context to the storage of memories about external events, objects and agents. The attention filter processes also make use of the state of Emotion:D-Emotion:A complexes to provide a semantic context for motivator filters. The quantitative emotion filters in figure 3 are set directly by the Attention:D mechanism. The intensity levels of these filters are set in response to the Emotion:D mechanisms and the reflective component of the emotion engine.

Learning in the emotion engine occurs in two ways. The reflective mechanism is being implemented using a recurrent neural network that reflects the CA hive communities. Training of the network is given in terms of preferred states within the overall emotional landscape of the cellular automata communities. Further work will look at other types of neural architectures for this and other parts of the emotion engine. The other learning mechanism is the development of preferred reactive behaviour (Emotion:R) combinations in the Emotion:D processes for a particular transition between the steady states of the Emotion:A communities. This is seen as an adaptation of the emotion engine in toto. Currently an experimental harness is being developed, using the Sim_Agent toolkit (Davis et al 1995), in which the emotion engine is trained to prefer specific combinations of emotions, for example the four emotions in similar valences (i.e. all negative, positive or neutral). Artificial scenarios are then provided in which the hive(s) are set in specific or random configurations. As different "personalities" prefer different aspects of the emotional landscape, the engine modifies itself so that preferred emotional states arise as valenced events occur, and preferred dispositions are maintained over longer time spans. Once satisfied that this framework is performing as expected, the earlier motivational architecture will be redesigned to incorporate the emotion engine. This will allow experimentation with emotionally-valenced motivators and allow the investigation of the referenced research using a deeper model of computational mind.

7. Future work

The primary reason for the *preliminary* research described above was to gain a better understanding of the relations between emotion, cognition and mind. Although earlier research on the computational modelling of motivation looked promising, there was a psychological implausibility with the motives behind motivators. Events in an agent's external environment can be represented in terms of motivational descriptors that connect the internal and external environments. The events in an agent's internal environments are described in terms of a synergy over different categories of (internal) computational processes that relates emotions, moods, personality and control. This paper places emotion at the core of the mind. This is analogous to the radioactive cores at the centre of a thermo-nuclear power plant. The plant needs those cores to function but they are not the full story to the functioning of the plant. If synthetic agents are going to experience emotions because of the nature of multiple-goal processing, then the computational infrastructure of those agents needs a representational framework in which these emergent qualities could be harnessed. The emotion engine is one small step in that direction.

While the described work may (superficially) satisfy Picard's (1997) five components for an agent experiencing emotions, the preliminary work is incomplete in a number of ways. The interplay of the reflective and reflexive components requires considerable more work. Preliminary experiments using MLP networks for the reflective processes proved unacceptable at the design stage. Current investigations look to mechanisms that move between discrete (three) space and the non-linear interval, with the queen-cells of currently active hives mirrored in the reflective network. This mechanism also needs to select the appropriate reactive (Emotion:R) behaviours for the preferred combination of emotional dispositions. A more sophisticated architecture would accept non-preferred emotional dispositions in order to achieve important (but temporally local) goals. Preferred dispositions are made non-extant while these goals are achieved. This is an issue that will need to wait until the emotion engine is placed within the architecture shown in figures 1 and 2. Then comparisons with other computational models of emotion, for example (Velásquez 1998) will be possible. Further analysis and investigation will determine whether it is possible to categorise emotion combinations in a manner analogous to the Ortony et al analysis. The discrete version of the basic set of emotions means there are at least 80 possible combinations of emotions; more if event, object and agent directed subtypes are considered. This paper has purposely ignored the social context for emotions, on which there is considerable study from Aristotle to today (see Elster 1999). This is a further inadequacy of the computational theory sketched here.

It has not been possible to review all pertinent evidence within the remit of this paper. The research into the nature of consciousness, and how it might be accomplished within a computational framework, has been glossed over. We accept Wollheim's differentiation between conscious, preconscious and unconscious mental states, and reiterate that any theory that underplays the role of emotions (and personality) in this and other mental phenomena is seriously flawed, as

suggested by over 100 years of neuroscientific, psychological and psychiatric evidence. Two of Wollheim's three levels of consciousness map onto the computational framework of reflexive, reactive, deliberative and reflective processes – the theory and model have yet to incorporate the unconscious. It remains unclear whether this will enable a computational agent to experience emotion in the same sense that biological agents experience emotion.

8. References

Beaudoin, L.P. and A. Sloman, A study of motive processing and attention, In: *Prospects for Artificial Intelligence*, Sloman, Hogg, Humphreys, Partridge and Ramsay (Editors), IOS Press, 1993.

Damasio, A.R. *Descartes' Error : Emotion, Reason and the Human Brain*, MacMillan Books, 1994.

Davis, D.N., Sloman, A. and Poli, R. *Simulating agents and their environments.* AISB Quarterly, 1995.

Davis, D.N., Reactive and motivational agents. In: *Intelligent Agents III*, J.P. Muller, M.J. Wooldridge & N.R. Jennings (Editors), Springer-Verlag, 1996.

Davis, D.N., T. Chalabi and B. Berbank-Green, Towards an architecture for artificial life agents: II, In: M. Mohammadian (Editor), *New Frontiers in Computational Intelligence and Its Applications*, ISO Press, 1999.

Davis, D.N., Modelling emotion in computational agents, *Paper submitted to ECAI2000*, 2000.

Ekman, P. and R.J. Davidson (Editors), *The Nature of Emotion*, Oxford University Press, 1994.

Elster, J., *Alchemies of the Mind: Rationality and the Emotions*, Cambridge University Press, 1999.

Frijda, N., *The Emotions*, Cambridge University Press 1986.

Goleman, D.P, *Emotional Intelligence*, Bloomsbury Publishing, 1995.

Gunderson, K., *Mentality and Machines* (2e), Croom Helm, 1985.

Harré, R., Emotion and memory: the second cognitive revolution. In: *Philosophy, Psychology and Psychiatry*, A.P. Griffiths (Editor), Cambridge University Press, 1994.

Hegselmann R. and Flache, A., Understanding complex social dynamics: A plea for cellular automata based modelling. *Journal of Artificial Societies and Social Simulation*, Vol. 1, No3, 1998.

Lazurus, R.S., *Emotion and Adaptation*, Oxford University Press, 1991.

Lindsay, P.H. and D.A. Norman, *Human Information Processing: An Introduction to Psychology*, Academic Press, 1972.

Newell A., *Unified Theories of Cognition*, Harvard University Press, 1990.

Norman, D.A., *Twelve Issues for Cognitive Science*. In: Issues in Cognitive Modeling, A.M. Aitkenhead and J.M. Slack (Editors), LEA Press, 1985.

Oatley, K. and Jenkins, J.M, *Understanding Emotions*, Blackwell, 1996.

Ortony, A., G.L. Clore and A. Collins, *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

Picard, R., *Affective Computing*, MIT Press, 1997.

Port R.F. and T. Van Gelder (Editors), *Mind As Motion*, MIT Press, 1995.

Power, M. and T. Dalgleish, *Cognition and Emotion: From Order to Disorder*, LEA Press, 1997.

Rolls, E.T., *The Brain and Emotion*, Oxford University Press, 1999.

Schneck, D., Music in human adaptation, CogNet Hot Science, http://cognet.mit.edu/, 2000.

Simon, H.A. Motivational and emotional controls of cognition, *Models of Thought*, Yale University Press, 1979.

Sloman, A. and M. Croucher, Why robots will have emotions. *Proceedings of IJCAI7*, 197-202, 1987.

Sloman, A., Motives, mechanisms and emotions, *Cognition and Emotion* 1, 1987.

Sloman, A. Architectural requirements for human-like agents both natural and artificial, In *Human Cognition and Social Agent Technology*, K. Dautenhahn, Benjamins Publishing, 1999.

Solomon, R.C., *The Passions*, Hackett, 1993.

Spaulding, W.D. (Editor), *Integrative Views of Motivation, Cognition and Emotion*, University of Nebraska Press, 1994.

Velásquez, J.D., When robots weep: emotional memories and decision-making, *Proceedings of AAAI-98*, 1998.

Wollheim, R., *On The Emotions*, Yale University Press, 1999.

Wright, I.P., Sloman, A. and Beaudoin, L., Towards a design-based analysis of emotional episodes, *Philosophy Psychiatry and Psychology, Volume 3*, 1996.