

# Emotion as the basis for computational autonomy in cognitive agents

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](mailto:D.N.Davis@dcs.hull.ac.uk)

**Abstract.** Many agent architectures are competency-based designs related to tasks in specific domains. More general frameworks map across tasks and domains. These types of agent architectures tend to fit well with the concept of weak notion of agency; i.e. they define autonomous systems that perform specific roles within a real or abstract environment. 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 – the foundational concepts underlying the concept of agency are no longer adequate for stronger notions of agency. Four foundations to the weak notion of agency are autonomy, social ability, reactivity and pro-activeness. These tend to be defined in terms of interactions between an agent's environment and the motivational qualities of an agent. From the perspective of developing intelligent computational systems this is more than acceptable. However, these definitions are shallow and insufficient for agent designs (and architectures) defined with regard to some aspect of cognitive functioning. There is no core to these agents other than an information processing architecture. 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 that underpins an agent's autonomy, social behaviour, reactivity and pro-activeness. As an agent functions it is sometimes called to monitor its internal interactions and relate the nature of these wholly internal functions to tasks in its external environment. The impetus for change within itself (i.e. to adapt or learn) is manifested as an unwanted combination (disequilibrium) of emotions. The modification of an agent's internal environment is described in terms of an emotion motivated mapping between its internal and external environments. 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*

## **Introduction**

At the first ATAL workshop, autonomy was defined as one of four foundations for a weak notion of agency [23]. 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". In the same volume, Castelfranchi [2] 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). Ferber [7] associates autonomy with "a set of tendencies, which can the form of individual goals to be achieved". This notion of autonomy is constrained by the concept of validity. An agent can only be autonomous within certain bounds. The boundaries to an agent's autonomy can be described in terms of niche spaces and design spaces. An agent inhabits specific (closely related) niches. An agent's niche space constrains the extent of its internal environment; the representations it is possible to use, what they relate to and how these representations form the basis for computational processes. An agent design space constrains what is possible for that agent. For example, a physical agent (or robot) may be equipped with auditory sensors but without the appropriate internal computational representations the niche space of music remains alien to it. This paper presents a perspective on autonomy using metaphors appropriate to stronger notions of agency. A stronger notion of agency relates to the modeling of more human-like qualities in cognitive agents. This perspective builds on current work on the emotions in cognitive science, neuroscience, philosophy and psychology. The impetus for this research is the inadequacy of earlier work on the modeling of motivation [5] to adequately contain aspects of agent functioning.

## **Why give agents emotions?**

Merleau-Ponty [10] supposes humans are moved to action by disequilibria between the self and the world. The impetus for thought and action in autonomous biological agents is a lack of cohesion in an agent's mapping and/or understanding of the relation between the agent's internal and external environments. In biological agents emotion is a primary source of motivation, and plays a large role in initiating and providing descriptors for these disequilibria. From a computational perspective, Sloman considers that intelligent machines will necessarily experience emotion (-like) states [18]. Following on from the work of Simon [17], his developing theory of mind and the nature of problem solving considers how perturbant (emotion-like) states

ensue in attempting to achieve multiple goals. Perturbant states will arise in any information processing infrastructure where there are insufficient resources to satisfy all current and prospective goals. This will occur not only at the deliberative belief and goal management levels but over all layers of the architecture as goals are mapped onto (internal or) external behaviors. An agent must be able to regulate these emotion-like states or compromise its autonomy.

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*” [21]. Schenck [16] 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*”. Emotions play an important role in the executive aspects of cognition, i.e. judgement, planning and social conduct. 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 that relate to current and important motivational interests. Such functions underpin notions of autonomy. Many researchers have written on the importance of emotion for motivation [22], memory [15], reason [3] and learning. In short emotion has a central role in a functioning mind. There is therefore a case for a computational model of emotion in the construction of agent theories and designs.

Thinking about, designing and building agent systems is guided by aspects of cognitive functioning. For example, a BDI agent architecture is a computational extension to human agents in thinking about problems using logic. The use of such rational models is understandable. They provide formal systems with well-defined properties. The limitations of such systems (e.g. logical omniscience) are known. Such systems are amenable to the reasoning of tasks and goals in the micro-worlds that most agents inhabit and internally represent. However, their use effectively circumscribes the depth of the foundational concepts underlying a currently adopted notion of agency. Where computational resources are constrained, goal-based agent models will necessarily have to prioritize which goals are to be pursued. If this leads to perturbant (emotion-like) states then these complex agents and agent societies need a computational model of emotion to manage these states or compromise their autonomy and reactivity. Even the most rational of agents will be emotionally compromised if it does not have the mechanisms to cope with the side-effects of determining which among many goals are those to be pursued.

An alternative stance is to place emotion at the core of agent processing. This provides an agent with a model of self that maps across different levels and types of processing. Emotion provides an internal source of autonomy and a means of valencing information processing events. In the remainder of this paper the influence of environment, both internal and external, upon this (agent-self) model of autonomy is considered. This emotional core gives rise to episodic states (e.g. feelings), trajectory states (e.g. moods and dispositions) and (semi-permanent) endogenous states (e.g. personality). These control states provide an agent with an internal model it can use to valence motivational aspects of its behavior. Through the development of

an appropriate computational model, an agent can regulate this basis for autonomy in terms of qualitatively different types of motivation.

### **Three psychological models of emotion**

Ortony et al [12] 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. 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. [1]).

Oatley and Jenkins [11] 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 (e.g. [8]) give similar definitions. A number of other psychologists (e.g. [14]) appear to be in agreement in defining the physiological, expressive and semantically distinct 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.

Rolls [15] presents a different perspective on the psychology of the emotions. Brains are designed around reward and punishment evaluation (or reinforcer) 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 behavioral patterns of responses, 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 (i.e. goals) are advanced and negative when impeded. These states are more encompassing than those states associated with the mere feelings of emotion. There is an overlap with the perspectives of Power and Dagleish, and Oatley and Jenkins. 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

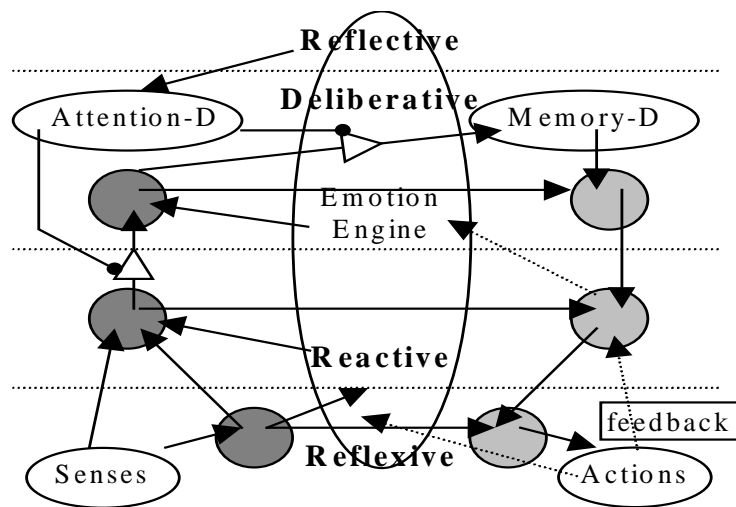
facilitation of memory processing (i.e. storage and control) and maintenance of “*persistent and continuing motivation and direction of behavior*”. These psychological models of emotion all relate to the four concepts underlying the weak notion of agency and are particularly relevant to the stronger notions of agency.

## **Theoretical framework**

Wollheim [24] 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 are more long-lived – 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 characterize 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 is 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 (mental states) associated with mental dispositions. While the two can be causally interrelated this need not always be the case. Mental dispositions are preconscious traits. In everyday functioning 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 the perspective of niche space and design space, we can use this (type of) analysis to model an agent and the roles and motivators (goals) it is expected to adopt in terms of its actual, possible and desired niche spaces.

Like Wollheim, Sloman differentiates between episodic and persistent mental phenomena, both of which can carry emotional constituents. His architectures for functioning minds include primary, secondary and tertiary emotions [20]. 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. Any agent architecture that supports multiple motivations (or goals) is subject to this type of dysfunction. In many situations these perturbant states arise through resource inadequacy or

mismanagement while pursuing multiple and not necessarily incompatible goals. Most agent implementations face this type of problem even if their underlying theory does not. The theoretical framework presented here revisits an earlier (computational) architecture of mind and emphasizes the interplay of cognition and emotion through appraisal, motivation and niche space. Emotions are in part mental (appraisal) states with supporting (valencing) and causal (reinforcer) processes. This provides a regulatory framework for the different forms of emotion inducing events. In moving towards a model of emotion that will be computationally tractable, the extent of the model will be initially (at least) minimized (ontological parsimony). A minimal model of emotion enables it to be used as the core to an agent-based model of mind.



**Fig. 1.** Sketch of the simplified four-layer architecture with emotion as the core. Dark gray circles represent information assimilation and synthesis processes. Light gray circles represent information generation processes that typically map onto internal and external behaviors

Earlier research on agents focussed on an architecture that supports motivation. The architecture (sketched in its simplest form in figure 1) emphasizes 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 has high level and low level processes co-existing and interacting in a holistic manner. The majority of the higher level processes tend to remain dormant, protected by filters and are activated only when sufficiently required. 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 logistics of goal satisfaction. Motivator processing, planning and other cognitive processes are not merely abstract but exist in relation to an agent's long term goals. An agent is autonomous to the extent that it determines how these long term goals (the reason for its existence) are to be achieved. The extent of its autonomy is governed by its design and the nature of its skills.

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 rationalization. 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 neighboring 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 behaviors) are possible for any such move. Some trajectories while impossible are supported or attended to for any number of reasons. Emotional intensity associated with a preferred niche space is one example. The possible trajectories between niche sub-spaces depend on an agent's design. The preferred trajectory between these niche spaces depends on autonomous preference for specific aspects of its emotional landscapes. An agent is autonomous to the extent that it can choose to pursue specific motivational trajectories. An agent is rational to the extent that it follows feasible (or achievable) trajectories. An agent's emotional landscape is the internal niche space that underpins an internally consistent understanding of external events, objects and agents – this is not always rational.

Emotions can be casually inter-related and cause other events. Drives and motivations are highly inter-linked with emotions. These can embody some representation and in effect relate short-term emotive states to temporally global processes. The control patterns that stabilize this model are the dispositions that influence the different categories of cognitive and animated behavior. An agent of a specific disposition will concentrate on certain tasks that favor specific aspects of the possible emotional landscape as external agents, objects and events affect emotionally valenced goals. Moods arise from the interaction of current temporally-global niche roles (the favoring of certain aspects of emotion space) and temporally-local drives that reflect the current focus of the agent. 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 that may be acted on. Emotional autonomy means an agent maintains an ongoing (globally-temporal) disposition with a range of (emotional) responses to specific events and their effect upon agent's overall emotional stance. The nature of this disposition is temporarily modified through current goals and motivations. Over time events occur that modify, stall, negate or satisfy goals. Such events can impinge on all layers of the architecture, affecting current dispositions and can lead to a recalibration of motivator preference. These events give rise to reinforcers. The emotion(s) they reinforce depends on their interactions with conscious and preconscious states and dispositions. Reinforcers and the (preconscious) valences can be modeled 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. Non-emotion low-level drives can be associated with reinforcers and be valenced. They can also be associated with motivators. The management and success (or otherwise) of these drive-generated motivations can give rise to emotions. 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 behaviors. This enables

emotions to be defined over different levels of the architecture using different aspects of motivational behaviors.

If emergent behaviors (related to emotions) are to be recognized 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 arising from anywhere in the architecture using explicit (motivator or goal) representations. The reflective processes classify the processing patterns of the agent in terms of combinations of the basic emotions and favored emotional dispositions. The emotion-changing (reactive) behaviors can be used to pursue a change in emotional disposition. Emotions can be instantiated by events both internal and external at a number of levels of abstraction, whether primary (e.g. ecological drives) or by events that require substantive cognitive processing. Emotions can be invoked through cognitive appraisal of agent, object or event related scenarios. The postponement (or abandonment) of a goal may cause an agent to experience emotion states related to unwanted dispositions. To move to a preferred aspect of the possible emotional landscape, an agent may need to instantiate other motivators and accept (perhaps) temporarily unwanted dispositions. An agent with emotional autonomy accepts temporary emotional perturbation if more acceptable niche spaces result. In the model in figure 2, 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 [5]. The deliberative appraisal of the emotion then activates laterally at the deliberative layer, affecting memory management, attention filters and motivator management. Changes to an agent's beliefs are possible through external influence, as in the Castelfranchi model of autonomy but here this process is mediated by the emotional autonomy of the agent.

## **Experimental computational work**

The architecture for a computational mind is based on ideas developed within the Cognition and Affect group at Birmingham [1,19,5]. 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 2 presents a four-layer processing model of the emotions. The ongoing automatic processes (Emotions:A) present a base for the model both for disposition processing and inflection of the ongoing dispositions through preconscious events. Such inflections are instantiated by events both external and internal to the agent. The reactive behaviors (Emotions:R) control the functioning of all the Emotions:A processes. The currently extant Emotion:R behaviors are set by deliberative processes (Emotions:D). The Emotions:M module encompasses the entirety of the meta-management (reflective) processes in this model of the mind. These reflective processes monitor the deliberative appraisal of the Emotions:A processes and the state of the attention filters (managed by Attention:D). The output from Emotions:M provides guidance to the attention management, Emotion:D and the Emotion:A processes. The agent attempts to learn to manage its emotions through the development of these five modules. Other aspects of the emotion engine are the



in response to the reflective processes. The deliberative processes also disturb the motivator management processes with their current emotive state – as part of an asynchronous appraisal mechanism. 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 emotional state of Emotion:D-Emotion:A complexes to provide a semantic context for motivator filters. The quantitative emotion filters in figure 2 are set directly by the Attention:D mechanism. The intensity levels of these filters are set in response to the Emotions: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. Training of the network is given in terms of preferred states within the overall emotional landscape of the cellular automata communities. The other learning mechanism is the development of preferred reactive behaviour (Emotion:R) selections in the Emotion:D processes for a particular transition between the steady states of the Emotion:A communities. Currently an experimental harness is being developed, using the Sim\_Agent toolkit [4], in which the emotion engine is trained to prefer specific combinations of emotions. Artificial scenarios are then provided in which the hive(s) are set in specific and random configurations. As the different agent “personalities” prefer different aspects of the emotional landscape, the engine modifies itself so that the preferred emotional state arises. 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 of varying transience in a number of domains.

## **Future work**

This 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 Boltzmann-like machines that move between discrete (three) space and the non-linear interval, with the center-cells of currently active hives mirrored in the reflective network. 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.

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 modeling of motivation looked promising, there was a psychological implausibility with the motives behind motivators. 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.

Through the development of stronger notions of agencies based on psychological models and philosophical analyses better-founded concepts underlying the weak notion of agency ensue. This paper has considered how even rational (logic-based) models of agents will be seriously undermined in terms of multiple-goal selection if the agent designs do not contain processes capable of harnessing perturbant internal behaviors resulting from such decision. The emotion engine is one small step in that direction.

## References

1. 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.
2. Castelfranchi, C. Guarantees for autonomy in cognitive agent architectures. In [23]: 56-70.
3. Damasio, A.R. *Descartes' Error: Emotion, Reason and the Human Brain*, MacMillan, 1994.
4. Davis, D.N., Sloman, A. and Poli, R. Simulating agents and their environments. *AISB Quarterly*, 1995.
5. Davis, D.N., Reactive and motivational agents. In: *Intelligent Agents III*, J.P. Muller, M.J. Wooldridge & N.R. Jennings (Editors), Springer-Verlag, 1996.
6. 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*, IOS Press, 1999.
7. Ferber, J. *Multi-Agent Systems*, Addison-Wesley, 1999
8. Frijda, N., *The Emotions*, Cambridge University Press 1986.
9. 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.
10. Merleau-Ponty, M., *The Structure of Behaviour*, Methuan:London, 1965.
11. Oatley, K. and Jenkins, J.M., *Understanding Emotions*, Blackwell, 1996.
12. Ortony, A., G.L. Clore and A. Collins, *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.
13. Picard, R. *Affective Computing*, MIT Press, 1997.
14. Power, M. and T. Dalgleish, *Cognition and Emotion*, LEA Press, 1997.
15. Rolls, E.T., *The Brain and Emotion*, Oxford University Press, 1999.
16. Schneck, D., Music in human adaptation, CogNet, <http://cognet.mit.edu/>, 2000.
17. Simon, H.A. *Models of Thought*, Yale University Press, 1979.
18. Sloman, A. and M. Croucher, Why robots will have emotions. *Proceedings of IJCAI7*, 197-202, 1987.
19. Sloman, A., Motives, mechanisms and emotions, *Cognition and Emotion* 1, 1987.
20. Sloman, A. Architectural requirements for human-like agents both natural and artificial, In *Human Cognition and Social Agent Technology*, K. Dautenhahn, Benjamins, 1999.
21. Solomon, R.C., *The Passions*, Hackett, 1993.
22. Spaulding, W.D. (Editor), *Integrative Views of Motivation, Cognition and Emotion*, University of Nebraska Press, 1994.
23. Wooldridge, M. and N.R. Jennings (Eds), *Intelligent Agents*. Springer-Verlag, 1995.
24. Wollheim, R., *On The Emotions*, Yale University Press, 1999.