

Agents, Emergence, Emotion and Representation

Darryl N. Davis

Neural, Emergent and Agent Technologies Group,
Department of Computer Science, The University of Hull,
Kingston-upon-Hull, HU6 7RX, UK

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

Abstract

This paper presents an analysis of issues pertaining to the relation between computational emergence and emotion in cognitive agent systems. We consider how a developing computational theory of cognition can be used to monitor and manage interactions with and within complex systems. Goal-based agent systems sometimes need to postpone or abandon important goals in favor of more immediate concerns. In more sophisticated goal-based agent systems goals may need to be abandoned due to conflict of motivations. This can lead to the emergence of perturbant behavior analogous in type to (negative) emotion-like states. Agent systems need mechanisms to recognize this type of scenario or otherwise risk compromising their design rationale. Through the development and use of a dynamic representation of emotion, a computational agent can harness unwanted and emergent states and behaviors before the agent system becomes dysfunctional. A comparison is made between emergent computational states and emotional states. This paper weaves these threads together in a consideration of the nature of sophisticated dynamic representations (e.g. goals and motivations), emotion and philosophical issues related to the modeling of cognitive agent systems and the relevance in future computational systems responsible for complex human-machine interactions.

Index terms-- agents, emergence, emotion, motivation

1. Introduction

Much of agent theory, design and technology centers on the management of goals (or goal-like states). An agent in being autonomous, reactive and pro-active responds to changes in its environment in such a way that it can pursue its (typically given) role or roles. As such an agent pursues the satisfaction of goals and motivations that maintain and further its role(s). Where conflict between goals arises, an

agent must autonomously determine which goals it can reasonably expect to achieve and postpone or abandon others. Much of agent-driven technology, while accepting that this is necessarily the case, places a shallow analysis on this postponement and abandonment of goals. Selecting between goals is typically described in the same terms as selecting rules in declarative (rule-based) knowledge based systems that make use of conflict resolution techniques. These types of agents (corresponding to the weak notion of agency) are shallow, differing from earlier artificial intelligence systems only in terms of their individual or social complexity. If we are to produce deeper (or broader) agents, with faculties that correspond to the types of cognitive processing found in socio-biological agents, these shallow models of goal resolution need to be treated with great circumscription. The postponement or abandonment of goals in socio-biological agents typically has emotive side-effects. The nature, quality and extent of the resulting emotion(s) depends on the agent, its (current) emotional propensity, the goal and what caused the postponement or abandonment of it. This paper considers why computational models of emotion are necessary in cognitive agents, and offers a developing computational framework that aims to achieve this. The conjecture that this paper supports is that the design and development of more robust complex systems will only be possible if we start to produce a more realistic cognitive agents that can interact with other agents, whether carbon or silicon based, in a more engaging and sophisticated manner. In short the socio-cultural matrix of homo-silicon interaction needs a silicon-based emotive content.

2. Need for Computational Models of Emotion

Merleau-Ponty [1] 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. An external environment is that the agent

perceives, communicates and acts upon, i.e. agents, events and objects external to an agent. An internal environment is one of perceptions, beliefs, knowledge and other dynamic processes and control states. In biological agents emotion plays a large role in initiating and providing descriptors for these disequilibria. For disequilibria to map onto effective actions in an agent's environment, these emergent states (and dispositions) need appropriate representational frameworks. Motivations (and goals) are examples of an appropriate representational form.

For human agents emotion is a primary source of motivation. Sloman [2] has for many years considered that intelligent machines will necessarily experience emotion (-like) states. Following on from the work of Simon [3], his developing theory of mind and the nature of motivation-based problem solving [4] considers how perturbant (emotion-like) states ensue in attempting to achieve multiple goals (or motivators). These perturbant states will arise in any information processing infrastructure where there are insufficient resources to satisfy all current and prospective goals (e.g. most current goal-based agent systems), or where important goals are mutually incompatible (e.g. HAL in the film 2001). 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*" [6]. Schenck [6] in his study of the role of music concurs with Sloman in suggesting 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*".

In terms of complex human-machine interaction, we need machines that recognize these aspects of cognition. Picard [8] discriminates between computational systems that can recognize affect and those that generate affective states. We suggest that for a machine to recognize a sequence of actions as inflected with emotional indicators, it needs to be more than an emotion recognizer. If a computational system is to reason about emotion, and causal events based on affective states, it needs to do more than recognize affective states in itself and others - it also needs to be able to generate these states. Emotion is by nature dynamic and to some extent emergent. Only through producing similarly dynamic representation of emotion (i.e. a processing model) can we design and implement machines capable of recognizing (and predicting) the affective nature of complex human-machine interactions.

3. Perspectives on Emotions

Research in psychology [9,10,11] is demonstrating how inter-related are cognition and emotion. Emotion has many functions including the valencing of emerging problems, tasks and challenges in terms of emotional intensity and

emotion type. The valencing of prospective tasks permits an agent to direct its attention to aspects of internal and external environments that relate to current and important motivational interests. Such a function is a precursor to problem solving as typically described in much AI research. Emotions play an important role in the executive aspects of cognition, i.e. judgement, planning and social conduct. Many researchers have written on the importance of emotion for motivation [3], memory, reason [10] and learning [11]. 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 of machines that are intended to display similar faculties.

Emotion can be described 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*" [9]. Hence, emotions are affective mental (conative and-or cognitive) states and processes. Conative states differ from cognitive states in being non-symbolic dynamic processes not readily represented using the control state space approach to mind [12]. These dynamic sub-cortical processes can be hormonally and chemically based. Such states do not however map easily into the representations used in symbolic reasoning. These categories of processes do, however, need to interface with the types of categorizations that can be modeled using cognitive and representational metaphors.

On the basis of physiologically, expressively and semantically distinct signs there is a case for considering five basic emotions [10]:

- ◆ 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

In fact, some form of continuum can be readily devised over an n-space matrix that includes axes for these five emotions with a further axis labeled from conative to cognitive. We could state that any particular emotional interaction could be sited on some level between a conative state to a cognitive state. Further valences (and axes) can be used to describe the current extent of the basic emotions, whether categorical (hate-dislike-like-love) or a more continuous label (for example normalized as a real number between -1 and 1). At any one time, the emotional state of an agent could then be represented, however simply, as some vector in this n-space.

This could then be used within the agent to categorize the nature of its current emotional imbalance (if any) and respond accordingly. The development of this as a computational metaphor within a dynamic cognitive architecture is problematic, but ultimately useful as it may provide for a more readily accessible environment for people to make more effective uses of interactional cyberspace.

Our stance in this work 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.

4. Emergence

Computational emergence can be categorized [13] in four ways. Diachronic emergence describes computational states that emerge over time, typically measured in terms of evolutionary periods. For example in applying evolutionary algorithms to design problems, the resulting optimized designs are a diachronic result emerging from the use of such systems. Gestalt emergence describes the processes whereby recognizable patterns of processing emerge, much like the interpretations that observers place on the patterns displayed by Conway's Game of Life. Representational emergence with which representational schemes related to behavior in at least one layer in a multi-layer computational system emerges at other layers. Representational emergence describes, for example, the results of the chunking mechanism in (symbolic) computational cognitive architectures such as SOAR [14]. Functional emergence describes situations whereby system capabilities emerge and roles are redefined through the redesign and runtime compilation of supporting mechanisms. Designers who rely on emergence for their computational systems typically make assumptions about how their systems respond to or make use of these different categories of emergence. These are described in terms of sub-cognitive bias, behavioral bias, and individualistic bias. However, given that we consider emotion to be, in part, an emergent property of complex agent systems, a more worrying category is emotional bias. Designers assume that their systems will respond to unexpected or adverse events in an emotionally neutral, rationale manner. With no framework to harness the emergent emotion-like behavior (a problem with all systems that are subject to emergent behavior) the analysis offered

by Sloman suggests that the emotional basis of the perturbant agent systems is likely to be negative; undermining the agent's role and expertise. Furthermore capitalisation upon harmonious situations is not possible.

Emotional bias can be defined as "*the likelihood of responding with one kind of emotion more than another*" [9]. In a homogeneous computational framework, to achieve specific goals certain very specific patterns of behavior are required of all users. However, because of the emotional response of individuals to any specific pattern of behavior, the emerging interaction may not necessarily be fruitful unless the computational framework has the capability to respond to the user's affective state. This problem is compounded where complex interactions with heterogeneous media and systems are necessary. Irrespective of the tasks modeled and whether they are internal to a complex system or the human interaction with such systems, perturbant computational behavior analogous to human emotive states can occur in these systems.

5. Theoretical Framework for Emotion Model

Wollheim [15] 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 (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 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 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 (mental states) associated with mental dispositions.

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 [16]. 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 liable 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 [12] and emphasizes the interplay of cognition and emotion through arousal, appraisal and motivation. In this framework, emotions are in part mental (appraisal) states and supporting (valencing) and causal (reinforcer) processes. This provides a regulatory framework for the different forms of emotion inducing events.

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 embody some representation (not necessarily semantic) 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 and 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 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 be instantiated by and give rise to more enduring motivational states that may be acted on. Over time events occur that modify, stall, negate or satisfy goals. Such events can impinge on 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. Non-emotion low-level drives are also permitted. These 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 given definitions of emotion is that they are described in terms of goals, roles and expressive behaviors. This enables emotions to be

defined throughout the architecture using different aspects of motivational behaviors.

6. Computational Model : Experimental Work

The architecture (sketched in its simplest form in figures 1 and 2) emphasizes four distinct processing layers: a reflexive layer analogous to the autonomic systems in biological agents, a reactive layer, a deliberative layer and a reflective layer. This broad picture has high and low level processes co-existing and interacting in a holistic manner. 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 and current processing across all layers of the architecture. An agent is autonomous to the extent that it determines how 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 environment related skills. If emergent behaviors (related to emotions) are to be recognized and managed in a computational agent then there must be a design synergy across the different structural and temporal layers of the agent architecture.

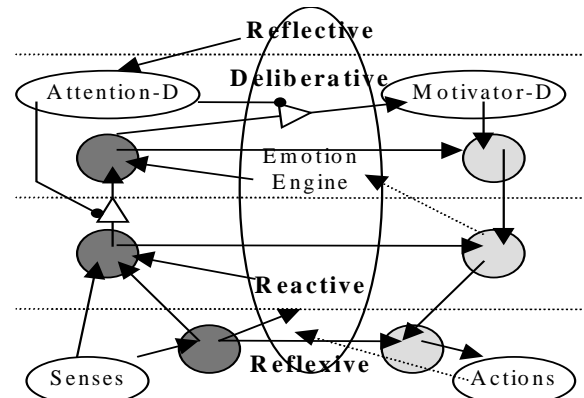


Figure 1. Sketch of the simplified four-layer architecture. Dark gray circles represent information assimilation and synthesis processes. Light gray circles represent information generation processes that typically mapping into internal and external behaviors

Processes at the deliberative level can reason about emergent states arising from anywhere in the architecture using explicit (motivator-related) representations. Reflective processes can classify the processing patterns of the agent in terms of combinations of the basic emotions and favored emotional dispositions. The emotion-changing behaviors can be used to pursue a change in emotional disposition. Aspects of emotions can be preconscious and emergent. Emotions can move into the conscious mind or be invoked at that level through the 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 (e.g. ecological drives) or by events that require substantive cognitive processing. In the model in figure 1, 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 [17]. The deliberative appraisal of the emotion then activates laterally at the deliberative layer, affecting memory management, attention filters and motivator management.

Figure 2 presents a four layer processing model of the emotions. The ongoing autonomic processes (Emotion: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 (Emotion:R) control the functioning of all the Emotion:A processes. The currently extant Emotion:R behaviors 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. These 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 subsequent Emotion:A processes. The agent attempts to learn to manage its emotions through the development of these five modules. All processes other than the Emotion:A module remain dormant (and filter protected) until aroused by external events or through autonomic processes elsewhere in the architecture.

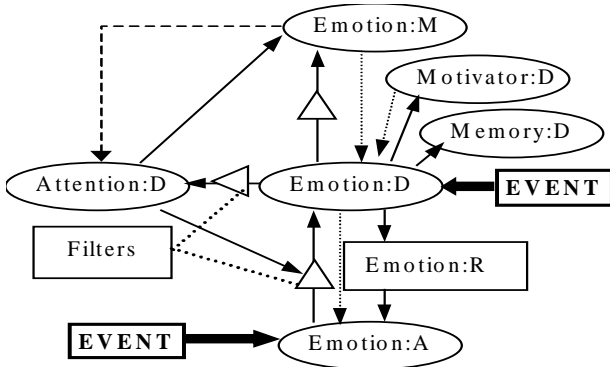


Figure 2. Sketch of the simplified four-layer emotion engine for figure 1.

For a number of reasons the Emotion:A module is modeled using multiple communities of cellular automata (or hives). The behaviors associated with the Emotion:R module are those that govern the internal behavior of single cells, the communication between adjoining cells in communities and inter-community communication. Four communities of three-state automata are used to represent four basic emotions (anger, disgust, fear and sobriety – the latter replaces both happiness and sadness). These form the automatic basis for the emotion engine, and ultimately underpin the notion of autonomy in the agents we are developing. Each emotion is discretely valenced in the CA communities as positive-neutral-negative. Further

community types represent reinforcers (valenced pre-emptive events). The behavior of each cell and inter-cell communication is governed by 4 sets of behaviors. Communication between different hives (and events outside of the emotion engine at the reflexive and reactive level) is by means of further (mobile) CA communities. The currently set behavior from these dispositions and disposition-event couplings is selected (at the reactive level) by deliberative (Emotion:D) processes responsible for asynchronously monitoring these communities in response to intense CA community states and to guidance from the meta-management (Emotion:M) module. Early experiments found that from a given state, the CA communities rapidly achieved one of a small number of generalised steady states, with over 30000 transitions possible. By changing the currently extant behavior set or by communicating with another hive (through the use of a mobile CA community) transitions to the same or other steady (sometimes oscillatory) states always occurred. Rules are used to select different CA community dispositions. Through the modification of the internal state of a small number of cells the emotion engine moves to a closely related state.

The deliberative processes change the agent's emotional preference (the temporally global emotive aspect of motivations), and hence the currently extant behavior set for the hives in response to the reflective processes. This in turn causes the deliberative processes to disturb the motivator management processes with their current emotive state – as part of the cognitive appraisal mechanism associated with emotional states. Memory management (Memory:D) also responds to the Emotion:D processes in order to provide emotional context to the storage and recall 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 1 are set directly by the attention processes. The intensity levels of these filters are set in response to the deliberative processes and the reflective component of the emotion engine.

Currently an experimental harness is being developed, in which the emotion engine will be 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 and random configurations. As different "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, earlier motivational architectures [12, 16] will be redesigned to incorporate the emotion engine. This will allow experimentation with emotionally-valenced motivators of varying transience in a number of domains.

7. Discussion

This is preliminary work and 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 stochastic 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 on further research.

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.

As cybernetic systems and the interactions with them become more complex they are in danger of disenfranchising their human creators. It is suggested that more harmonious trajectories through the landscape of human-machine interaction will result through a consideration of goal-based processing that takes account of emotional involvement. Here we conjecture that a more appropriate emotion engaging perspective on goals can only be provided through a consideration of the niche spaces and emergent properties of emotional behavior. Cybernetic systems are being used to solve problems, but most current metaphors and descriptions of problem solving involve the idea of (quite constrained) directed thinking. De Bono [18], for example, rails against the rigidity of conceptual delineation, considering it to hinder creative thought and problem solving. Creative problem solving is sometimes linked with the emotions. Perhaps interactions that include affective components will lead to more creative solutions. We need further research and analysis on the nature of emergence and emotion to understand how to design and develop more effective systems, irrespective of the underlying computational stance of their creators. While this

paper provides an incomplete analysis of the cognitive environment appropriate from an engaging and meaningful interaction within the socio-cultural-homo-machine-ecology, it is suggested that such like analyses ought to be considered when developing complex systems, whether they are to be used for entertainment, education, business or medicine.

References

- [1] M. Merleau-Ponty, *The Structure of Behavior*, Methuen, London, 1965.
- [2] A. Sloman and M. Croucher, Why robots will have emotions. *Proceedings of IJCAI7*, 197-202, 1987.
- [3] H.A. Simon, *Models of Thought*, Yale University Press, 1979.
- [4] L.P. Beaudoin 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.
- [5] A. Sloman, Motives, mechanisms and emotions, *Cognition and Emotion* 1, 1987.
- [6] R.C. Solomon, *The Passions*, Hackett, 1993.
- [7] D. Schneck, Music in human adaptation, CogNet Hot Science, <http://cognet.mit.edu/>, 2000.
- [8] R. Picard, *Affective Computing*, MIT Press, 1997.
- [9] K. Oatley and Jenkins, J.M., *Understanding Emotions*, Blackwell, 1996.
- [10] M. Power and T. Dalgleish, *Cognition and Emotion*, LEA Press, 1997.
- [11] E.T. Rolls, *The Brain and Emotion*, Oxford University Press, 1999.
- [12] D.N. Davis, Control states and complete agent architectures, *Computational Intelligence*, 17, 2001.
- [13] N. Gilbert and R. Conte, *Artificial Societies: The computer simulation of social life*, UCL Press, 1995.
- [14] A. Newell, *Unified Theories of Cognition*. Harvard University Press, 1990.
- [15] R. Wollheim, *On The Emotions*, Yale University Press, 1999.
- [16] A. Sloman, Architectural requirements for human-like agents both natural and artificial, In *Human Cognition and Social Agent Technology*, K. Dautenhahn (Editor), Benjamins Publishing, 1999.
- [17] D.N. Davis, Reactive and motivational agents. In: *Intelligent Agents III*, J.P. Muller, M.J. Wooldridge & N.R. Jennings (Editors), Springer-Verlag, 1996.
- [18] E. de Bono, *Water Logic: The Alternative to I am Right You are Wrong*. Viking Press, 1993.