Proceedings of MASHO 02, German Conference on AI *KI 2002*, Karlsruhe, Germany 2002

# Enhancing Perception and Planning of Software Agents with Emotion and Acquired Hierarchical Categories

Joscha Bach

Humboldt University of Berlin
Department of Computer Science, Artificial Intelligence
bach@informatik.hu-berlin.de

**Abstract.** The implementation of situated agents that mimic aspects of human or animal cognition requires psychological theories with regard to motivation, perception, emotion and memory that are both detailed and formal. The 'Psi' theory of psychologist D. Dörner provides a framework for agents that fulfills some of these demands and focuses on emotional modulation of perception, action-selection, planning and memory access. This paper is an attempt at giving a short introduction to artificial emotion, some aspects of Dörner's theory and briefly hints at possible lines of extension.

## 1 Introduction

Perception and planning in complex environments are challenging and fascinating tasks to implement within software agents. As soon as the domain of micro-worlds is left, agents need to take measures to keep the resulting complexity of their percepts and the respective ontologies in check. In the following, two ways of achieving this are briefly outlined: using 'emotional states' to reduce the complexity of searches, and using acquired hierarchical categories. The author is currently implementing an agent structure loosely based on the 'Psi'-theory of emotion by Dietrich Dörner [Dörner 1999], which is enhanced by hierarchical categories.

### 1.1 Emotions in 'Psi' agents

While there are numerous approaches to emotional concepts for software agents, the Psi-theory is unique in that emotions are not defined as explicit states but rather emerge from modulation of the information processing and action selection. Thus, they are not explicit building blocks of the agents but exist 'in the eye of the beholder', as agglomerated descriptions of mental states of agents, or, to be more specific, as dispositions to action, perception and planning. What does this mean?

Dörner's agents react to their environment by forming memories, expectations and immediate evaluations. Additionally, they possess a number of fixed, but individually different parameters that influence their behavior and perception. Because the resulting behavior is sufficiently complex, the task of describing and predicting it in terms of a complete analysis of internal states becomes difficult, if not intractable.

Nonetheless, external observers find that their behavior yields to intuitive explanation: when agents are in grave danger, they do not longer examine their environment closely, but try for obvious escapes. Depending on the internal self-assessment of the agents, they seem to approach solutions with vigor or fright. Contrariwise, an agent that is 'healthy' and 'well fed' may display serendipitous behavior or curiosity. Because it is difficult to distinguish internal states with more detail within expressionless agents, Dörner's agents indicate their reactions also by a number of graphical displays, including a face that is animated in accordance to the respective theory of emotion. Thus, it becomes possible to attribute mental states (specifically, emotional episodes) to the agent that allow for plausible (albeit limited) explanation and prediction of its behavior. While the agents implement Dörner's theory of emotions, the emotions themselves have not been explicitly defined within the agents; they have not been programmed to act *as if* they had emotions. Rather, they become apparent because the agents reflect their interaction with the environment in certain parameters, and the *resulting agent configurations* and their changes resemble emotional episodes in biological agents. This notion of emotions as a class of mental states that can be meaningfully attributed on the behavior of a situated agent is close to what Daniel Dennett has called the *intentional stance* [Dennet 1971], and to Aaron Sloman's concept of attributed *virtual architectures* [Sloman 1994].[1]

This approach also has drawbacks for some applications: because the emotional state of the agents is compositional and heterogenous, there is no simple mapping to sets of 'top level' basic emotions as proposed in widely used emotion models (for instance OCC [Ortoni, Clore, Collins 1988]). This makes it more difficult (but not impossible) to describe Psi's emotions in terms of these categories.

### 1.2 Artificial emotion

While the term 'emotion' is often used rather freely in agent design, it is by no means completely clear what emotions in humans, or, more generally, organisms are, and consequently, there are a lot of ways to define artificial emotions.

Let me briefly approach the term 'artificial emotion'. Like for its 'natural' counterpart, there are many dissimilar notions, which focus on different aspects, typically depending on the *function* authors attribute to emotion, or on the *application* within the project.

A major application might be social simulation. In many scenarios, models of humans as rational agents are insufficient. Thus, whenever concepts of *bounded rationality* are regarded, a need to form non-rational models of agent behavior arises. An example for this is Bernd Schmidt's agent framework PECS [Schmidt 2001]. PECS is a three-layered architecture with a controlling middle layer harboring specialized rule-based modules to represent physiological, emotional, cognitive and social states. PECS is not a theory of emotion or cognitive behavior, rather, it is meant

[1] Note that Dennet and Sloman disagree on the ontological status of the ascribed mental states. In very loose terms it might be said that Dennet considers these mental states to be convenient shorthands to whatever is going on, while Sloman puts them in line with more directly observable empirical phenomena.

to allow for a simplified implementation of the behavioral aspects relevant to the simulation.

Another, quite obvious field of application is the implementation of *believable agents* like computer-animated actors in movies or computer games. Here, the modeling of emotional states is a much more important prerequisite to the creation of credible characters than that of rational behavior [Reilly 1997]. Believable agents are not only vital for entertainment purposes, but are also already in use for interfaces to information systems and online shop front-ends. While such interface agents do not actually have to undergo emotional states but merely have to display them, interaction with human users requires a theoretical understanding of emotions that is sufficient for the establishment of credible communicative exchange.

Similar models may aid in the building of systems that process textual documents: For instance, when it comes to the automatic translation of novels, the accurate modeling of the mental states of the described human actors may be crucial to find equivalent expressions in different languages.

Furthermore, there is a scientific interest to find better models of emotion and test them in simulations. An example of the latter is *Cathexis* [Velasquez 1997, 1999], an architecture that has been developed with regard to neurophysiological findings. The main focus of this paper is a work in the same line that is concerned with arriving at an understanding of emotion that allows the software agents to undergo emotional states. Clearly, this is linked to a particular notion of biological emotion and a certain understanding of the role of emotions in cognitive processes, which will not stand undisputed. While many researchers tend to agree that emotions are a prerequisite for many cognitive capabilities (for example, [Minski 1986] and [Damasio 1994]), or at least an important aid, some argue that emotions are mainly a hindrance to efficient rational behavior and appear as 'perturbances' in the flow of cognition [Sloman 1992].

However, this is not necessarily a contradiction, because the label 'emotion' is often glued on very different phenomena.

### 1.3 Human emotions

There is no general agreement on the role or even the nature of emotions in humans; neither in common usage nor in the scientific literature. I.e., authors disagree on whether to subsume or exclude hedonic aspects ('feelings'), situation-based evaluations, motivations, endocrinal configurations, facial expression feedbacks etc. There is also no agreement as to which classes of mental phenomena should be termed 'emotion' (like affects, moods, emotional dispositions and so on). The discussion if the different views of the relationship between emotions and drives, concerns, situation assessments and physiological effects is way beyond the scope of this paper. It is important to note, however, that emotions are not identical with their physiological correlates, but are effects on the information processing. In other words: emotions are intertwined with dispositions to perceive, imagine, recall, memorize, plan, and act *in a certain way.* If we say, that someone 'has an emotion', we imply that this individual is inclined to certain ways of perception, may prefer certain kinds of action or may plan in a different way, and so on. In the understanding of Dörner's

'Psi' theory, these implications *are* what makes up these emotions, i.e. if the system has sufficient means of perception, interaction, planning etc. and can modify them to adapt to changing situations, then this system is effectively undergoing (but not necessarily experiencing) emotional states.

Of course, such a system would not automatically implement the full range of human emotions, because many of them require cognitive faculties that have not been implemented in software agents yet. Emotions can be seen as integral parts of cognitive systems in such a way that they underlie other layers of behavior control. This means that they can only modulate behavior the agent is already capable of, for instance, social emotions can not occur in agents that are unable to perceive, identify and model other agents. Furthermore, these emotions require certain basic dispositions to be built into the agent by which it is driven to relate to its environment in a certain way.

## 2 The Psi model

Within the 'Psi' model of emotion, Dörner claims that for the emergence of emotional states, the modulation of cognitive processes according to internal and external demands can be sufficient, that is, these modulations lead into states that would be (in conjunction with beliefs and desires) perceived as emotional by external observers, and by the agent itself, if it has self-reflecting capabilities. Within the framework of this theory, several modifiers have been proposed, like 'resolution level', 'selection threshold' and 'activation'. Together with built-in motivators, representing desires for resources, intactness, competence, reduction of uncertainty and affiliation, they produce indeed complex behavior that can be interpreted as being emotional, and reproduces the behavior of human actors in the same situation (i.e. in the same simulated agent-world) to a considerable extent [Dörner 2002].[2]

### 2.1 Perception and internal representation

The information processing of the Psi agents is based on the idea that the elements of perception and imagination are fundamentally the same. Thus, perceptions are imaginations which are inspired by and verified against external sensory data. The basic building blocks of perception and cognition are linked structures of feature-representing nodes with varying activation, called 'Schema'. (In a way, these structures are similar to 'cases' in Case Based Reasoning.) The links between nodes represent causal or membership/part relations, can be enhanced with temporal and spatial data and can be hierarchically organized. The schema for a flower object in the agent world might consist of the possible parts of a flower (along with their spatial relationship to each other), and will be linked to actions that have been learned to be

[2] In experiments featuring an island world, which the agent could explore in search for food, water, mayhem and so-called nucleotides (a bonus item), the behavioral patterns of different classes of human subjects setting out on the same endeavor could be successfully mimicked by choosing appropriate parameters for planning dispositions, competency estimates etc.

possible or impossible to perform with flowers, to results that are to be expected of flowers, to precursors of flowers, and to contexts in which flowers have been perceived so far. Each part of the flower is termed a sub-schema and may consist of further sub-schemas (currently, line-elements are the lowest element of perception).

### 2.2 Actions and planning

Actions and transitions are practically triplets of schemas – they consist of a 'before' and an 'after' situation, and the 'motoric' action that has to be issued in-between. (Again, this action may contain different sub-schemas, which 'bottom out' in commands to the agent's actuators.) It is relatively simple to perform planning with these triplets; after a goal is chosen, the agent has to find a chain of matching triplets leading from the current world state to the goal.

Nonetheless, the complexity of this search can be enormous, and is reduced by several means:

- *Context*. Given objects and motivations raise the activation of 'connected' schemas and direct perception, memory retrieval and so on accordingly. (*"If you have a hammer, everything starts to look like a nail. Especially nails."*) This is implemented by a spreading activation approach, where schemas pre-activate related schemas on which more intimate searches take place.

- *Modification of the search*. By traversing the memory-graphs in more depth, more width, with higher activation (larger context) or narrower focus (faster, more straightforward results) different trade-offs may be chosen according to the situation at hand. These modifications are achieved with the modifiers mentioned above, and lead to 'emotional configurations'.

- *Using different search algorithms*. The evaluation of different ways of finding a chain of actions (like searching forward, backward, several directions simultaneously, prefer hill-climbing or accept temporary disadvantages) may lead to the preference of strategies according to situation, or the discovery of new strategies. This is currently not implemented.[3]

Obviously, in situations of reasonable complexity, plans can only achieve a length of very few steps, because the search space limits the results. Thus, the modification of search strategies according to the situation ('emotional modulation') becomes crucial; for example, in dangerous situations, a bias towards quick, obvious solutions might be beneficial. This can be achieved by reducing the depth and selection threshold of the search. If creative solutions are in demand, the search strategy should be able to

[3] Eventually, search strategies should be expressed as schemas themselves and be subject to meta-deliberation by the agent. This requires the schemas to be connected to symbolic reasoning mechanisms. Current work by Dörner's group is concerned with the organization of schemas by a simple language including a spacial calculus.

follow non-obvious connections, and pursue side-tracks to a considerable depth ('serendipitous' search).

### 2.4 Acquisition of categories

While these configurations improve the capabilities of the agent considerably, a major improvement is to be expected by cognitive enhancements:

- *Hierarchies*. By finding appropriate super-schemas for action sequences (like: 'eat' instead of 'moving your hand towards the edible object, close your fingers around it, check whether you managed to get hold of it, move it to your mouth, insert it, close your mouth, chew etc.'), plans can be extremely reduced in size. The same holds true for individual objects (like 'hand' instead of 'finger tips, nails, joints, palm etc.'). Only when it comes to performing the actions, individual sub-actions need to be expanded. These strategies are only implemented to a small degree and still subject of research.

The problem is the finding of appropriate super-schemata. Currently, the Psi-agents do this by grouping frequent protocol chains, and by implicitly classifying objects that 'look' or interact likewise. It is important to note that things like 'hand' or 'eat' are not 'directly observable entities', but categories superimposed on perception and cognition. Some of these categories are easy to infer by statistical means, others may be less obvious and require a lot of 'categorical experiments' before they can be established and stabilized. (See for instance the establishment of the notion of 'force' in individual cognitive development [Ioannides, Vosniadou 2002]).

For the formation of more complex categories, the use of at least some aspects of language may be a prerequisite for two reasons: First, many categories can possibly only be built by connecting concepts derived from highly abstract notions, and second, mechanisms of communicative exchange between agents will facilitate the fast exchange of the results of a large agent population, on which can be built from thereon. (For experiments on category-formation in communicating agents, see for instance Luc Steels [2001]).

## 3 Some links to related work in AI

The agents of Dörner's research group have been designed primarily with respect to research in theoretical psychology. While their design shows numerous influences from AI, few similarities and differences to related work have been discussed by the original authors. In part, Dörner's agents represent re-inventions of concepts that have been developed in different areas of AI, which leads to a somewhat different terminology. There are also few attempts of the Dörner group at formalizing their concepts and architecture. The establishments of links into technically related work may prove very fruitful for the future development of the Dörner project.

The basic building block of 'sensor schemas' is the 'Quad', which consists of a central node (typically representing some feature) and four auxiliary nodes, which provide links to other quads. Of the resulting four classes of links two are causal ('ret' for backwards and 'por' for forwards causation), and two represent class/member or whole/part relations ('sur' for links to parts, and 'sub' for the opposite direction). The nodes are linked in such a way that central nodes are connected to their auxiliaries only, and auxiliaries are connected with a number of matching auxiliaries from other quads (i.e. 'ret' to 'por' and 'sur' to 'sub'). The strengths of the links are determined by (positive or negative) weights. Furthermore, each node has an activation value and a threshold. Activations are computed by summing the weighted activations of input nodes and are cut when below the threshold. In addition to these links, activations may be set using general activators (these will activate one of the four link-directions throughout the complete net, thus controlling the spread of activation). According to Dörner, the four classes of links correlate to the Aristotelian four classes of *causae*.

With these links, information can be organized in hierarchies. This does not only apply to percepts, but also to actions: frequently occurring chains of action schemata are chunked into a single schema that links onto them with a sur-sub connection (much like the chunking in SOAR, for instance. [Laird, Newell, Rosenbloom 1987]).
By using the ret-por arcs as conjunctions, the sur-sub arcs as disjunctions and schemas as predicates, it is possible to express statements in an *n*-order logic.

The links between quads are established and deleted by an additional kind of node input, called 'associator' and 'dissociator'. Whenever the associator of a node is active, it is linked to all other currently active nodes. Vice versa, by activation of the dissociator, it is possible to remove or weaken links to other, currently active nodes. Dörner also utilizes chained quads as scripts for the immediate control of the agent.
Because the quad net is also capable of Hebbian learning, Dörner calls it a neural network. Although this is not incorrect, it might be more accurate to term it a hierarchical causal network or belief network [Good 1961] [Shachter 1986], because the nodes typically represent features or cases (see [Russel/Norvig 1995]). (Belief networks are often also called Bayesian networks, influence diagrams or knowledge maps.)

Nevertheless, there are some differences to typical implementations of belief networks, namely, there are no pre-defined inheritance mechanisms, and differences between part and membership relations are missing. Links on the same level of hierarchy are expressed by making indistinct use of causal (ret/por) arcs.

The mechanisms of memory building and retrieval allow for an associative memory (see for instance [Anderson 1973]).

The organization of memory is also very similar to *Case Retrieval Networks* (CRN) [Burkhard 1995] from the domain of Case Based Reasoning. Here, sensory nodes are equivalent to *Information Entities* (IEs) and schemas are called *cases*. If schemas are hierarchical, intermediate schemas are equivalent to *concept nodes*. Horizontal links (ret/por connections) are somewhat comparable to *similarity arcs*, while vertical links are akin to *relevance arcs*. Especially during memory retrieval, the analogy to CRNs with spreading activation (SAN) becomes obvious. Again, there are some differences: similarity arcs in CRNs are undirected, which often leads to problems if the activation is spread more than one or two steps due to loops in the linking of the IEs. [Lenz 1997]. On the other hand, ret/por connections do not really

depict similarity but relate elements at the same level of a belief network, so the process of spreading activation does not only find new, similar super-schemas (cases) but activates more details of the currently activated structures.

Another similarity that springs to mind is that to the *Copycat* architecture [Hofstadter 1995] [Mitchell 1993]. Unlike Copycat, the long term memory of the Psi agents is not strictly subdivided and holds episodic data, object data (without explicitly distinguishing between instances and classes) and action sequences. But Dörner's causal networks show parallels to what is called the *slip net* in Copycat. Like the slip net, the quad net uses spreading activation and allows for analogy making in hierarchical concepts by way of allowing conceptual 'slippages' (switches between similar sub-schemas). However, this is not very pronounced in current implementations of Psi. Copycat does not include an emotional model, nor are action selection and retrieval modulated in the same way as they are in Dörner's agents. (However, there is a modulator, called *temperature*, that influences the level on the concept hierarchy on which conceptual slippages are likely to occur.)

*IDA* and *CMattie* [Franklin 1999] are more recent agent designs that combine Copycat's slip nets with models of emotion, but here emotions do not modify the retrieval. Instead, they mainly act as additional properties of retrieved concepts and help to determine context and relevance.

## 4 Conclusion

The Psi theory and their current implementation as situated agents unite work from several areas of AI in a unique way and as such provide an inspiring influence both for the design of agents and understanding the relationship between emotion and cognition. They also include interesting models of perception, imagination and planning (most of which have not been described with sufficient detail here). To take them beyond that and enable the utilization and evaluation of the Psi concepts, the author believes that a more formal model must be derived. This is the object of current work of the author's workgroup, along with the design of mechanisms for category and hierarchy building.

**References:**


Anderson, J., Bower, G. (1973). *Human Associative Memory*. Washington, DC: Winston

Burkhard, H. D. (1995). *Case Retrieval Nets.* Techn. report, Humboldt University, Berlin

Damasio, A. R. (1994). *Descartes' Error*. *Emotion, Reason and the Human Brain*. Avon Books

Dörner, D. (1999). *Bauplan für eine Seele*. Reinbeck: Rowohlt

Dörner, D. (2002). *Psi. Eine neuronale Theorie der menschlichen Handlungsregulation*. (in print)

Dennet, D. (1971). *Intentional Systems.* The Journal of Philosophy, 68(4):82-106

Franklin, S. (1999). *Action Selection and Language Generation in "Conscious" Software Agents.* Proc. Workshop on Behavior Planning for Life-Like Characters and Avatars, i3 Spring Days '99, Sitges, Spain

Good, I. J. (1961). *A Causal Calculus*. British Journal of the Philosophy of Science, 11:305-318

Hofstadter, D. R. (1995). *Fluid Concepts and Creative Analogies.* Basic Books, New York

Ioannides, C., Vosniadou, S. (2002) *The Changing Meanings of Force*. in: Cognitive Science Quarterly, ed. Strube, Vol. 2, Iss. 1

Laird, J. E., Newell, A., Rosenbloom, P. S. (1987). *SOAR: An Architecture for General Intelligence.* Artificial Intelligence, 33: 1-64

Lenz, M. (1997). *CRNs and Spreading Activation.* Technical report, Humboldt University, Berlin

Minsky, M. (1986). *The Society of Mind*. New York: Simon and Schuster.

Mitchell, M. (1993). *Analogy Making as Perception.* Cambridge, MA: MIT Press

Ortony, A., Clore, G. L., Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge, England: Cambridge University Press

Reilly, S. (1997). *Believable Agents*. PhD Thesis, School of Computer Science, Carnegie Mellon University

Russel, S. J., Norvig, P. (1995). *Artificial Intelligence. A Modern Approach.* Prentice Hall, New Jersey

Shachter, R. D. (1986). *Evaluating influence diagrams*. Operations Research, 34:871-882

Schmidt, B. (2000). *PECS. Die Modellierung menschlichen Verhaltens*. SCS-Europe Publishing House, Ghent

Sloman, A. (1992). *Towards an information processing theory of emotions*. http://www.cs.bham.ac.uk/~axs/cog_affect/Aaron.Sloman_IP.Emotion.Theory.ps.gz

Sloman, A. (1994). *Semantics in an intelligent control system*. Philosophical Transactions of the Royal Society: Physical Sciences and Engineering. Vol 349, 1689, 43-58

Steels, L. (2001). *The talking heads experiment.* Obrist, H.U. and Vanderlinden, B. (eds) *Laboratorium*, 413-419, Cologne: DuMont

Velásquez, J. (1997). *Modeling Emotions and Other Motivations in Synthetic Agents.* In: Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97). Providence, RI: MIT/AAAI Press.

Velásquez, J. (1999). From affect programs to higher cognitive emotions: An emotion-based control approach. http://www.ai.mit.edu/people/jvelas/ebaa99/velasquez-ebaa99.pdf