Emotional Computers

Computer models of emotions and their significance for Emotion Psychology Research

Gerd Ruebenstrunk

Content

List of figures and tables

1. Introduction

What role do computer models play in emotion psychology research? This question is not new, but a number of answers are. Since the literature review by Pfeifer (1988), there have not been significantly more, but much more profound approaches that enable a more comprehensive answer to the original question.

In his essay "Artificial Intelligence Models of Emotion", Pfeifer (1988) not only tried to give a comprehensive overview of the existing modelling approaches, but also classified the existing models. For him, the approaches can be divided into two categories:

a) Augmented Cognitive Models
Approaches in this category do not focus on emotions, but consist mainly of models for cognitive processes in which emotions play a "complementary" role. Typical of such models is that they deal with a well-defined task in which emotions are added as additional factors.

b) AI Models of Emotion
Approaches in this category focus on modeling emotions. Typical for such models is the basic assumption of a complex environment in which clear task descriptions are difficult to realize.

In another work, Pfeifer (1994) modified his classification. Now he distinguishes between "reasoners" and "psychological models". *Reasoners* are models that are based on specific taxonomies of emotions and have the task of classifying emotions. Psychological models are systems whose goal is to model emotional processes per se.

The following work is primarily interested in what Pfeifer calls *reasoners* , but also wants to include approaches that cannot be integrated into any of the categories mentioned. It therefore assumes a different structure. Computer models of emotions are differentiated according to two different objectives:

a) Computers that can "understand" and "express" emotions;
b) Computers that "possess" emotions.

The second group is particularly interesting for emotion psychology research, although it raises most epistemological problems. While the models in the first category consist only of the practical implementation of more or less formulated theories of emotion and thus represent more of a technical challenge, the development of computers that possess emotions is about initiating an evolutionary process that ideally leads to the independent emergence of an emotional subsystem.

This thesis will therefore focus on approaches of the second category and present them in more detail in their historical and theoretical context. Technical explanations that are necessary for understanding the implementation of the models are dealt with as briefly as possible.

After an introductory overview (Chapter 2), the epistemological dimension of computational modeling of emotions is first discussed (Chapter 3). Then, in a brief overview, the psychological theories of emotion that serve as the basis for computer models of emotions are presented (Chapter 4). This is followed by a presentation of the most important models from the first category (Chapter 5). The main part of the thesis deals with models of the second category. The work of Simon (chapter 6) and Toda (chapter 7) is presented in more detail, as well as some first implementations of Toda's model (chapter 8). As a large-scale implementation, a chapter is dedicated to Sloman's model (chapter 9) as well as its continuation by Wright (chapter 10). Toda's approach in 1998 has led to a plethora of approaches to constructing emotional autonomic agents, some of which are briefly introduced (Chapter 11). In a final assessment (chapter 12), the significance of the models described in this thesis for emotion psychology research is examined.

2. Artificial feelings

The most famous emotional computer of all time is probably HAL 9000 from the movie "2001 - A Space Odyssey". If the vision of an artificial intelligence equal to, if not superior to humans, was already a shock for many moviegoers, it was even more frightening that this machine even had feelings - and therefore ultimately destroyed the people on board.

It was probably no coincidence that one of director Stanley Kubrick's advisors was Marvin Minsky, one of the fathers of artificial intelligence. For him, the "emotional computer" is a thoroughly realistic version:

> “ ... I don’t think you can make AI without subgoals, and emotion is crucial for setting and changing subgoals. Kubrick probably put the emotion in to make good cinema, but it also happens to be very good science. For instance, HAL explains that the Jupiter mission is too important to be jeopardized by humans. It is through emotion that he sets the goals and subgoals, ultimately killing the humans...”
> (Stork, 1997, p. 29)

In the meantime, the representatives of artificial intelligence have also recognized how important emotions are for an "intelligent" computer. This insight has grown less from a deep reflection on the topic, but rather from the failure of classical AI. The new buzzword is no longer Artificial Intelligence, but AE, i.e. Artificial Emotions.

The idea of an emotional computer, like the idea of an intelligent computer, is more of a threat than a hopeful vision for most people. On the other hand, such an idea exudes a strange fascination. It is not for nothing that intelligent and emotional machines play a not insignificant role in popular culture.

In addition to the aforementioned HAL, for example, there is the *Terminator*, who, in the film "Terminator 2" by James Cameron, is a robot without any feelings, but in the course of the film learns to understand human feelings. In one scene it even looks as if he is experiencing an emotion himself - whether this is really the case, however, the director leaves us in the dark. Another example is the robot from "Number 5 Lives", which transforms from a war machine into a good "human".

Robots, at least in popular culture, are described as alien beings whose character is predominantly threatening. They share this characteristic with the "real" aliens. Just think of Mr. Spock from "Star Trek", whose mind knows only logic, but no emotions - like all the inhabitants of his home planet Vulcan. And yet, in the course of the series, we are repeatedly reminded that he can't do without feelings either.

And even the monster "Alien" from the film series of the same name frightens us humans with its malicious intelligence, but, as at least the last of the four episodes makes clear, is not immune to rudimentary feelings.

It could therefore be concluded that an alien intelligence actually only becomes existentially threatening to us humans if it has at least a minimum of emotions. Because if it consisted only of pure logic, its behavior would be predictable and ultimately controllable by humans.

So it is no wonder that emotions have now also found their way into artificial intelligence. The Affective Computing research group at MIT explains the need to develop emotional computers as follows:

> “The importance of this follows from the work of Damasio and others who have studied patients who essentially do not have "enough emotions" and consequently suffer from impaired rational decision making. The nature of their impairment is oddly similar to that of today's boolean decision-making machines, and of AI's brittle expert systems. Recent findings indicate now that in humans, emotions are essential for flexible and rational decision making. Our hypothesis is that they will also be essential for machines to have flexible and rational decision making, as well as truly creative thought and a variety of other human-like cognitive capabilities.”
> (Affective Computing Home Page)

Although Damasio's works are of a more recent nature, the position represented here is not new, but can be traced back to the sixties. However, it has been forgotten again, at least in artificial intelligence. However, the inability of computers to perform complex actions with a high degree of autonomy has revived interest in this approach.

While the focus of AI research in the past was more on the representation of knowledge stocks, today it is concentrated on the development of "intelligent autonomous agents".

The interest in autonomous agents results from a number of practical requirements, e.g. from space travel. For example, it is desirable to use robots to explore distant planets that can perform their tasks there independently, as external control would be difficult to realize, e.g. due to the great distance. Another category of autonomous agents are software agents, which can perform a variety of different activities. The focus is on information collection tasks, for example on the Internet.

Franklin and Graesser define an autonomous agent as follows:

> “An **autonomous agent** is a system situated within and part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.”
> (Franklin and Graesser, 1996, p. 4)

Picard formulates the area of application for "emotional" autonomous agents a little more specifically:

> “One of the areas in which computer emotions are of primary interest is software agents, computer programs that are personalized - they know the user’s interests, habits and preferences - and that take an active role in assisting the user with work and information overload. They may also be personified, and play a role in leisure activities. One agent may act like an office assistant to help you process mail; another may take the form of an animated creature to play with a child.”
> (Picard, 1997, S. 193f.)

According to these definitions, autonomous agents can be implemented as a software-only solution – a limitation that is rejected by a number of researchers. For example, Brustoloni (1991) defines an autonomous agent as a system capable of autonomous, purposeful action in a real world, which is an appropriate and timely response to external stimuli.

Pfeifer (1996) also considers physical implementation to be an indispensable prerequisite for an autonomous agent, especially if he is to have emotions. His four basic principles for a "real life" agent according to the "fungus eater principle" (see below) are:

a) autonomy
The agent must be able to function without human intervention, supervision or instruction.

b) self-sufficiency
The agent must be able to keep itself functional over a longer period of time, i.e. it must be able to conserve or replenish its energy resources, maintain or repair its technical functions, etc.

c) embodiment
The agent must be able to have a physical body through which to interact with the physical world. This body is particularly important:

> "Although simulation studies can be extremely helpful in designing agents, building them physically leads to surprising new insights... Physical realization often facilitates solutions which might seem hard if considered only in an information processing context."
> (Pfeifer, 1996, S. 6)

d) situatedness
The agent must be able to control his entire interaction with the environment himself and be able to bring his own experiences into his interaction with the environment.

A taxonomy for autonomous agents, as presented by Franklin and Graesser (1996), makes it clear that autonomous agents of all kinds are fundamentally different from humans:

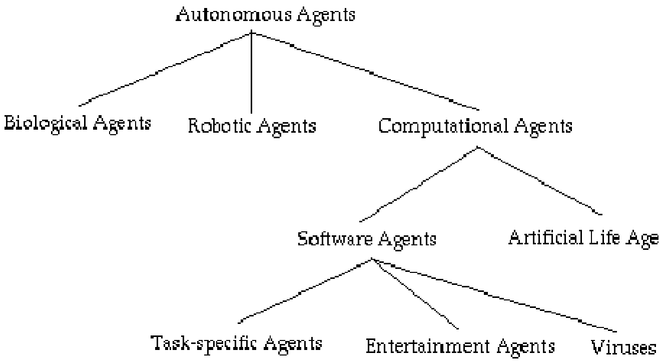


Fig. 1: Taxonomy of autonomous agents (Franklin and Graessner, 1996, p. 7)

It is the demand for a physical implementation that has brought together the fields of robotics and artificial intelligence. Within the framework of this cooperation, individual aspects of the above-mentioned principles have already been implemented; however, there is no implementation of a complete system that would meet all these requirements.

Despite the increased interest in autonomous agents, attempts to create intelligent machines must all be considered failures so far, even if, for example, Simon (1996) claims otherwise. Franklin (1995) very vividly outlines the three main AI debates of the last 40 years, each of which has grown out of the failure of the previous approaches.

There is no doubt that each of these failures has driven the development of intelligent machines - but, according to Picard (1997), a significant part is still missing. And that part is the feelings.

It is interesting that the growing interest in emotions in AI research has a parallel with the growing interest in emotions in cognitive psychology. While emotion theorists have tended to lead a somewhat marginal existence in recent decades, they have also been experiencing an appreciation in recent years. This has certainly also been due to findings in neuroscience (see e.g. LeDoux, 1996), which ascribe a much higher importance to the emotional subsystem for the functioning of the human mind than previously assumed.

Another parallel can be observed in the growing interest in the topic of "consciousness". This discussion has also been brought into psychology primarily from the circles of artificial intelligence, neuroscience and philosophy. However, even a cursory glance through some of the major publications shows that the old dichotomy between cognition and emotion continues here: Almost none of the present works on the subject of *consciousness* deals with emotions.

It is undisputed that at least a number of emotions cannot exist without consciousness. In particular, these are all the emotions that presuppose a conception of the "self", for example shame. Without wanting to go into more detail about the discussion of "primary" and "secondary" emotions here, it can be stated that there are emotions that occur independently of consciousness; but also emotions that presuppose consciousness.

In summary, it can be said that currently the impulses for the construction of real emotional systems come mainly from three disciplines: robotics, artificial intelligence and cognitive science. "Classical" research in the psychology of emotions has so far been rather cautious to wait-and-see, both in the reception of research contributions from these areas and in the delivery of its own contributions.

3. Strange Brains

Ever since computers have existed, there have also been attempts to simulate processes of human thinking on them. The ability of an electronic machine to read, manipulate and output information at high speed led researchers to speculate about the equivalence of computers and brains from the very beginning.

Such speculations soon found their way into psychology. In particular, the ability of computing machines to process information in parallel corresponded to approaches in psychology that viewed the brain primarily as a *parallel processing system* .

Against this background, computers were seen as a way to elucidate as yet unexplored phenomena of the human mind through modeling. A good example of this is Oliver Selfridge's Pandemonium model (Selfridge, 1959). Pandemonium is a model for visual pattern recognition. It consists of a large number of demons working in parallel, each of which specializes in recognizing a specific visual stimulus, for example, a crossbar in the center of the presented stimulus or a curvature in the upper right corner.

When a demon recognizes "its" stimulus, it calls out a corresponding message to the central *master-demon*. This call is all the louder the higher the demon assesses the probability of correct cognition. All demons work independently of each other; no one is influenced by his neighbors.

Based on the information received, the *master-demon* then decides which pattern constitutes the stimulus. In a further development of the model, the demons were organized hierarchically to relieve the *master-demon* .

There is a striking similarity between these specialized demons in Selfridge's model and the actual *feature detector* cells in the visual cortex. And indeed, it was Selfridge's model and the assumptions it made about perceptual processes that first suggested the idea that such *feature detectors* could exist in humans. In this case, the model was the reason for the neurophysiologists to search for corresponding cells.

This makes Pandemonium a good example of how computer models can advance psychological research. On the other hand, however, it should not be concealed that a system like Pandemonium is incapable of actually "seeing". And this is where the criticism ties in, which grants computer modeling a limited heuristic use, but otherwise denies any equivalence between man and machine.

Also and especially in the development of emotional computers, one of the fundamental questions is of course the equivalence of the systems "human" and "computer". Both AI research and previous approaches to the development of emotional systems assume without further inquiry that "intelligence" and "emotion" in a computer are not fundamentally different from intelligence and emotion in humans.

This assumption is largely abstracted from the specific hardware in each case; "Emotion" is also understood as a pure software implementation. It is quite questionable whether the two systems obey the same laws of development.

A computer is a discrete system that initially knows nothing more than just two different states. The combination of several such elements can cause "intermediate states"; however, this is only possible in a certain level of abstraction from the underlying hardware.

In contrast, the physiology of the emotional and cognitive systems in humans is by no means comparable mechanics, but consists, even at the lowest level, of a multitude of mechanisms, some of which work more according to digital principles, others more according to analogue principles.

Even one of the best-studied mechanisms, the function of neurons, is not exclusively an on/off mechanism, but consists of a multitude of differentiated sub-mechanics - and, as I said, at the hardware level.

The simulation of such mechanisms is currently only possible on computers as software. It is true that simple neuronal switching patterns can also be modeled hardware-wise by parallel computers up to a certain level; however, such modelling is only possible for a narrowly limited area and also completely ignores chemical processes, which also play an important role in humans.

Picard (1997) tries to solve the problem by abstracting from the difference between hardware and software and always understanding "computer" to be both. She justifies this position with the fact that emotional software agents can certainly exist in an "emotion-free" hardware.

A similar discussion is also being held about the comparability of emotions in humans and animals (see Dawkins, 1993). This is at least based on hardware made up of identical elements, albeit with varying degrees of complexity. In this case, too, it is not yet scientifically decided whether an emotion such as "grief" is congruent in humans and animals.

The matter is further complicated by the question of whether a computer can in principle be a life form. This question has received some attention in the "Artificial Life" discussion in recent years. For example, the evolutionary biologist Richard Dawkins (Dawkins, 1988) is of the opinion that the ability to reproduce alone would be sufficient to be able to speak of life in the biological sense. Others expand the definition to include the components "self-organization" and "autonomy".
If one disregards the discussion about "life", which is mainly conducted from an ethical and philosophical point of view, and concentrates on the aspects of "self-organization" and "autonomy", then it is quite realistic to ascribe these properties to computers or software. Self-organization in the sense of adaptation has been demonstrably present, for example, in neural networks that operate with genetic algorithms (see e.g. Holland, 1998). Autonomy in the limited sense has been proven in robots or, to some extent, in autonomously operating programs, for example agents for the Internet. Such programs also have the ability to reproduce, which would also meet the third condition.

The focus of AI and AL research is currently on the further development of such autonomous, self-organizing systems. The models used are partly based on functional models of the human brain; however, this should not be too quick to equate their functioning with that of the human brain.

For example, especially in the case of software optimization processes using genetic algorithms, human observers are often not aware of the self-organization processes that software uses to achieve the optimization goal.

Apart from such complex software processes, it may also make sense to attribute mental abilities to a computer when running simple programs. John McCarthy, one of the pioneers of artificial intelligence, explains:

> ".. Although we may know the program, its state at a given moment is usually not directly observable, and the facts we can obtain about its current state may be more readily expressed by ascribing certain beliefs and goals than in any other way... Ascribing beliefs may allow deriving general statements about the program's behavior that could not be obtained from any finite number of simulations. The beliefs and goal structures we ascribe to the program may be easier to understand

> than the details of the program as expressed in its listing... The difference between this program and another actual or hypothetical program may be best expressed as a difference in belief structure."
> (McCarthy, 1990, S. 96)

The attribution of mental information according to these explanations has only a functional nature: it serves to express information about the state of a machine at a given time that would otherwise only be expressible by lengthy and complex detailed descriptions.

McCarthy names a number of mental characteristics that can be attributed to computers under these aspects: introspection and self-knowledge, consciousness and self-awareness, language and thought, intentions, free will, understanding and creativity. At the same time, however, he also warns against equating such ascribed mental qualities with human characteristics:

> "The mental qualities of present machines are not the same as ours. While we will probably be able, in the future, to make machines with mental qualities more like our own, we'll probably never want to deal with a computer that loses its temper, or an automatic teller that falls in love? Computers will end up with the psychology that is convenient to their designers..."
> (McCarthy, 1990, S. 185f.)

In the meantime, we know that the last sentence does not necessarily have to be true. For there are first examples of self-organizing and optimizing hardware (Harvey and Thompson, 1997), the functions of which are not known to their human designer. And the current approaches to the design of emotional computers go far beyond modelling and actually try to develop computers whose mental qualities are not predetermined by the designer, but are intended to be created independently.

Although certain basic assumptions of the designers still flow into such projects, this approach is fundamentally different from the classical modelling that has been observed in cognitive science so far. The question, however, is whether the processes in a computer, which may one day actually develop mental qualities as a result of this approach, are identical to the processes in the human body and brain.

Critics of such approaches like to argue that emotions are not comparable to purely cognitive processes, since they are influenced by a number of additional factors (e.g. hormones) and also require a subject. The modelling of these processes is therefore not possible in a computer, a purely cognitive entity, above all because a machine lacks the subjective element of feeling an emotion whose essential component is a feeling. There are several answers to this argument.

On the one hand, it is not excluded in principle that a computer can have a subjective feeling. Computers are an extremely young phenomenon in evolutionary history and have undergone an astonishing development in this short time. Today, there are computers with hundreds of processors working in parallel; work is being done worldwide on the development of far more powerful bio- and quantum computers. Therefore, it is only a matter of time before a computer can have a similar complexity as the human brain in terms of hardware. As complexity grows, so does the probability that such a system will self-organize at a higher level. What is still laboriously programmed as a "monitor instance" today may well develop into something that could be described as the "I" of a computer in the foreseeable future.

Secondly, it would be anthropocentric to deny an intelligent system that does not have a human hormonal balance the ability to express emotions. Numerous "physiological" processes take place in a computer, which, equipped with appropriate proprioception, it could well perceive as "body feeling". If, in addition, it is an adaptive and mobile computer, it would be quite conceivable that it reacts to certain situations with a noticeable change in processes that have the same significance for it as physiological changes in our body.

An emotional computer does not necessarily have to experience emotions like a human, nor does a visitor to Zeta Epsilon. Nevertheless, his emotions can be as real for him as they are for us - and influence his thoughts and actions just as much as they do for us.

So we cannot assume a priori that "emotions" that develop in a computer are comparable to human emotions. However, it is quite plausible that the emotions of a computer fulfil the same functions for it as for us humans. If this is the case, computer modeling of emotions would not only be a step towards learning more about the meaning of emotions for us, but also creating the basis for a time in which intelligent systems of different "designs" will cooperate with each other.

4. Theoretical basics

In this part, a brief overview of the psychological theories of emotion that underlie the computer models presented below will be given. The claim here cannot be to present and critically evaluate every theoretical approach in all its complexity. Rather, it is a matter of presenting some core elements of the respective theoretical approaches, insofar as they are taken up again in the computer models.

It is interesting to note that the majority of computer models of emotions, insofar as they explicitly refer to psychological theory, are based on the so-called assessment theories.

The fascination of these approaches apparently lies in the fact that they are excellently suited for operationalization and are therefore (relatively) easy to implement in program code.

4.1. The theory of Ortony, Clore and Collins

Ortony, Clore and Collins (1988) have developed their theoretical approach explicitly with a view to implementation in a computer:

> "..., we would like to lay the foundation for a computationally tractable model of emotion. In other words, we would like an account of emotion that could in principle be used in an Artificial Intelligence (AI) system that would, for example, be able to reason about emotion."
> (Ortony, Clore und Collins, 1988, S. 2)

The theory of Ortony, Clore and Collins assumes that emotions arise as a result of certain cognitions and interpretations. Therefore, it focuses exclusively on the cognitive triggers of emotions.

The authors postulate that three aspects determine these cognitions: events, agents, and *objects*.

> "When one focuses on events one does so because one is interested in their consequences, when one focuses on agents, one does so because of their actions, and when one focuses on objects, one is interested in certain aspects or imputed properties of them *qua* objects."
> (Ortony, Clore und Collins, 1988, S. 18)

Emotions, according to their central assumption, represent valenced *reactions* to these perspectives on the world. One can be pleased/displeased about the consequences of an event; you can approve/*disapprove* of an agent's actions, or you can like/dislike aspects of an object.

Another differentiation is that events can have consequences for others or for oneself, and that one acting agent can be another or oneself. The consequences of one event for another can be divided into *desirable/undesirable*; the consequences for oneself as relevant or irrelevant expectations. Finally, relevant expectations for oneself can be differentiated according to whether they actually occur or not (*confirmed/disconfirmed*).

This differentiation results in the following structure of emotion types:

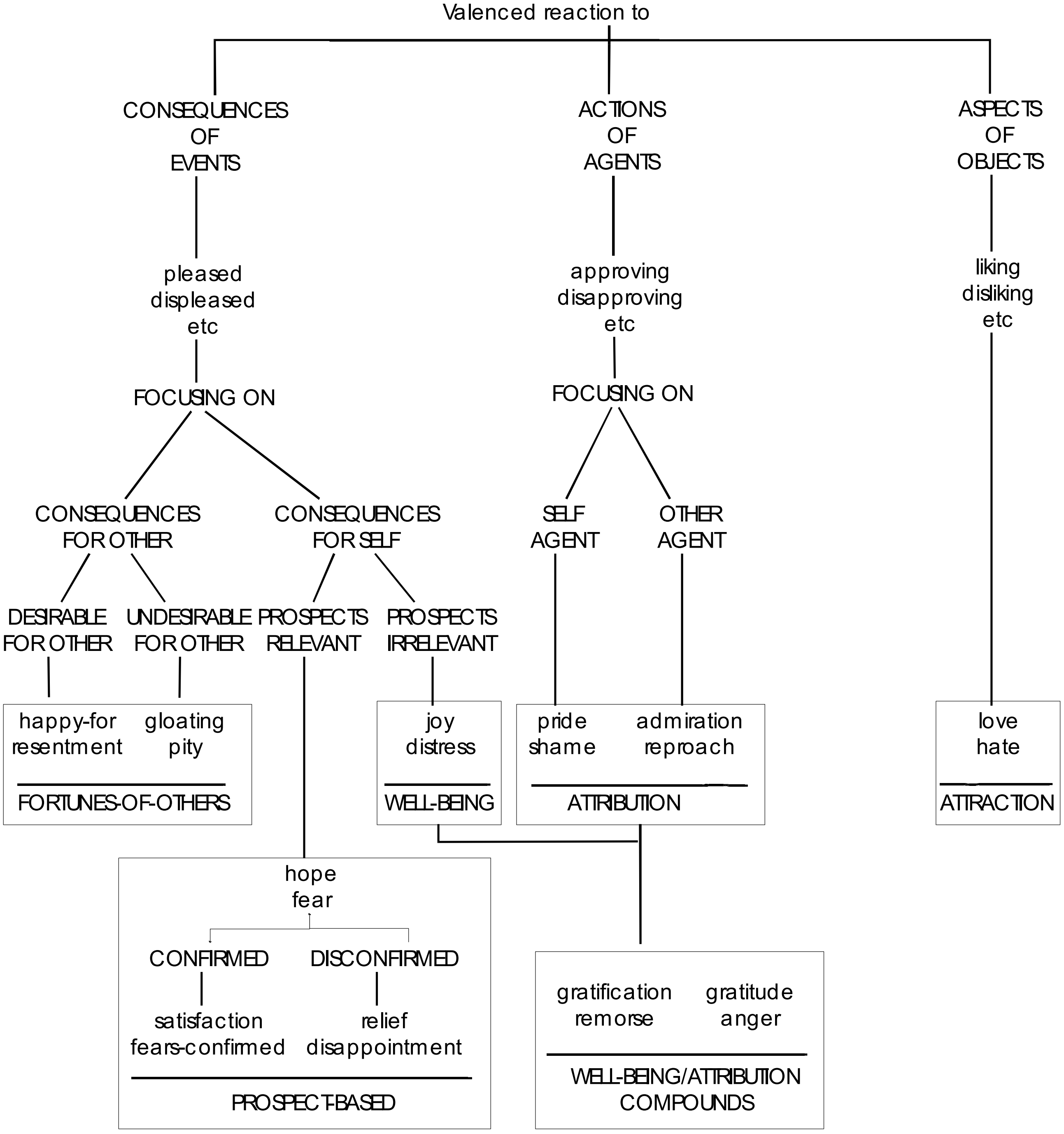


Fig.2: Structure of emotion types in the theory of Ortony, Clore and Collins (according to Ortony, Clore, Collins, 1988, p.19)

The intensity of an emotional sensation is mainly determined by three central intensity variables: *Desirability* is linked to the reaction to events and is evaluated with regard to *goals*. *Praiseworthiness* is linked to the reaction to actions of agents and is evaluated with regard to standards. *Appealingness* is linked to the reaction to objects and is evaluated with regard to attitudes.

Furthermore, the authors define a number of global and local intensity variables. *Sense-of-reality*, *proximity*, *unexpectedness* and *arousal* are the four global variables that act across all three emotion categories. The local variables, which include the central intensity variables mentioned above, are:

EVENTS	AGENTS	OBJECTS
desirability	*praiseworthiness*	*appealingness*
desirability for other	*strength of cognitive unit*	*familiarity*
deservingness	*expectation deviation*	
liking		
likelihood		
effort		
realization		

Table 1: Local variables in the theory of Ortony, Clore and Collins (according to Ortony, Clore and Collins, 1988, p. 68ff.)

Each of these variables is assigned a value and a weighting in the specific case. In addition, there is a threshold value for each emotion, below which an emotion is not subjectively perceived.

On the basis of this model, the emergence of an emotion can be described in formal language: *Let* $D(p,e,t)$ be the desirability (D) of an event (e) for a person (p) at a certain point in time (t). This function has a positive value for a desirable event and a negative value for an undesirable event. Furthermore, $Ig(p,e,t)$ is a combination of global intensity variables and $Pj(p,e,t)$ is the potential for a state of *joy*. Then the following rule can be created for "joy":

```
IF      $D(p,e,t) > 0$
THEN    set $Pj(p,e,t) = fj(D(p,e,t), Ig(p,e,t))$
```

The resulting function *fj* triggers another rule that determines the intensity for joy (Ij) and thus triggers the experience of joy-emotion. Let Tj be a threshold value, then the following applies:

```
IF      $Pj(p,e,t) > I.e.(p,t)$
THEN    set $Ij(p,e,t) = Pj(p,e,t) - I.e.(p,t)$
ELSE    set $Ij(p,e,t) = 0$
```

If the threshold is exceeded, this rule produces the emotion of joy; otherwise, it gives the value of "zero", i.e., no emotional sensation. Depending on the intensity of the emotion, different *tokens* are used to describe it. Such *tokens* are words that describe this emotion.

Ortony, Clore and Collins themselves do not provide formalizations for all the emotions they define, but only give a few examples. However, they claim that every emotion can be formulated in appropriate formal notation, even if this is far more complex for many emotions than for the example presented.

With the help of such a formal system, a computer should be able to draw conclusions about emotional episodes that are presented to it. The authors deliberately limit their goal:

> “Our interest in emotion in the context of AI is not an interest in questions such as “Can computers feel?” or “Can computers have emotions?” There are those who think that such questions can be answered in the affirmative…, however, our view is that the *subjective experience* of emotion is central, and we do not consider it

> possible for computers to experience anything until and unless they are conscious. Our suspicion is that machines are simply not the kinds of things that can be conscious. However, our skepticism over the possibility of machines having emotions certainly does not mean that we think the topic of emotions is irrelevant for AI..... There are many AI endeavors in which the ability to understand and reason about emotions or aspects of emotions could be important."
> (Ortony, Clore und Collins, 1988, S. 182)

4.2. Roseman's theory

Roseman's theory, which he first presented in 1979 (Roseman, 1979), was modified by him several times in the years that followed. In the process, it has changed in (in some cases essential) details; only the basic approach of an *appraisal theory* of emotions has remained the same.

Roseman developed his first theory based on 200 written accounts of emotional experiences. From the analysis of these documents, he derived his model, in which five cognitive dimensions determine whether an emotion occurs and which one.

The first dimension assesses whether a person has a motivation towards a desired situational state or a motivation away from an undesirable situational state. The dimension thus knows the states "positive" and "negative".

The second dimension covers whether the situation corresponds to the person's motivational state or not. The dimension thus knows the states "situation exists" or "situation does not exist" (*present* and *absent*).

The third dimension is whether an event is perceived as certain or merely as a possibility. This dimension knows the states "safe" and "unsafe".

The fourth dimension records whether a person perceives the event as deserved or undeserved, with the two states of "deserved" and "undeserved".

Finally, the fifth dimension captures from whom the event originates. This dimension knows the states of "the circumstances", "others" or "self".

From the combination of these five dimensions and their characteristics, a table can be compiled (Roseman, 1984), from which, according to Roseman, emotions can be predicted.

A total of 48 combinations of Roseman's dimensions can be formed (positive/negative x present/absent x safe/uncertain x deserved/undeserved x circumstances/other/self). According to Roseman, 13 emotions correspond to these 48 cognitive assessments.

After experimental tests of this approach did not yield the results postulated by Roseman, he modified his model (Roseman, 1984). The second dimension of his original model (situation present or absent) was now given the characteristics "motive-consistent" and "motive-inconsistent", whereby "motive-consistent" always corresponds to the "positive" expression of the first dimension and "motive-inconsistent" to the "negative" expression of the first dimension. The alternatives "present" and "absent" are now replaced by the terms "appetitive" and "aversive".

Another correction concerned the fourth dimension of the original model (deserved/undeserved). Roseman replaced it with the dimension of strength, i.e. whether a person feels strong or weak in a situation. The characteristics of this dimension are then also "strong" and "weak".

Roseman also added another expression to the third dimension of his original model (safe/unsafe): "unknown". This was necessary to capture the emotion of surprise in his model.

Roseman himself admits (Roseman et al., 1996) that this model has also proven to be insufficient in empirical testing. As a consequence, he developed a third version of his theory (Roseman et al., 1996). It differs from his second approach in several respects: The fourth dimension (strong/weak) is replaced by a relational assessment of one's own control potential, with the characteristics "low" and "high". The expression "unknown" of the third dimension is replaced by the expression "unexpected", as this is the prerequisite for the emotion of surprise, according to Roseman. And finally, Roseman adds another dimension to the negative emotions, which he calls the "problem type". It records whether an event is perceived as negative because it blocks a goal (resulting in "frustration") or because it is inherently negative (resulting in "disgust"). This dimension knows the characteristics "non-characteristic" and "characteristic".

The extent to which this last model variant of Roseman for the time being can be empirically proven cannot be assessed at present. However, one weakness of the model is evident: it has problems with a person making two different assessments in one and the same situation. If, for example, a student thinks that his teacher is administering an unfair test, but at the same time knows that he has prepared inadequately for the test, then it is not clear from the Roseman model what his emotions are - because there are two states of the fifth dimension in parallel.

Due to its simple structure, which can be excellently implemented in rules that define exactly according to which assessments which emotion occurs, Roseman's models have received a positive reception in artificial intelligence circles. For example, Dyer based his model BORIS on Roseman's first approach, and Picard also writes: "Overall, it shows promise for implementation in a computer, for both reasoning about emotion generation, and for generating emotions based on cognitive appraisals." (Picard, 1997, p. 209)

4.3. Scherer's theory

For Scherer, five functionally defined subsystems are involved in emotional processes. An information-processing subsystem evaluates the stimulus through perception, memory, prediction and evaluation of available information. A supportive subsystem regulates the internal state by controlling neuroendocrine, somatic and autonomic states. A leading subsystem plans, prepares actions and selects between competing motifs. An acting subsystem controls motor expression and visible behavior. Finally, a monitor subsystem controls the attention paid to the current states and relays the resulting feedback to the other subsystems.

For Scherer, the information-processing subsystem is of particular interest. According to his theory, this subsystem is based on assessments that Scherer calls *stimulus evaluation checks* (SEC). The result of these SECs in turn triggers changes in the other subsystems.

Scherer sees five major SECs, four of which have *subchecks*. The *novelty check* decides whether external or internal stimuli have changed; its subchecks are suddenness, familiarity and predictability. The *intrinsic pleasantness check* determines whether the stimulus is pleasant or unpleasant and causes corresponding approaches or avoidance tendencies. The *goal significance check* decides whether the event supports or hinders the person's goals; its subchecks are goal relevance, outcome probability, expectation, supportive character, and urgency. The *coping potential check* determines the extent to which the person believes he or she is in control of events; his subchecks are agent, motive, control, power, and adaptability. Finally, the *compatibility check* compares the event with internal and external standards and norms; its subchecks are externality and internality.

According to Scherer, each emotion can be clearly determined by a combination of the SECs and subchecks. A corresponding table with such classifications can be found in [Scherer, 1988]. A number of empirical studies have supported Scherer's model so far.

4.4. Frijda's theory

Frijda points out that the word "emotion" does not refer to a "natural class" and that it is not able to refer to a well-defined class of phenomena that are clearly distinguishable from other mental and behavioral events. For him, therefore, the process of emotion formation is of greater interest.

At the heart of Frijda's theory is the concept of *concern*. One concern is the disposition of a system to prefer certain states of the environment and of one's own organism to the absence of such states. Concerns produce goals and preferences for a system. When the system has problems realizing these concerns, emotions arise. The strength of an emotion is essentially determined by the strength of the relevant concern(s).

Frijda defines six essential characteristics of the emotional system that describe its function:

1. *Concern relevance detection*: The emotion subsystem reports the importance of events for the concerns of the overall system to all other components of the system. Frijda calls this signal affect. To do this, it must be possible to perceive information from the environment and from one's own system.
2. Appraisal: Next, the importance of the stimulus for the concerns of the system is assessed. This is a two-step process with the sub-processes relevance *appraisal* and *context appraisal*.

3. Control *precedence*: If the relevance signal is sufficiently strong, it changes the priorities of perception, attention and processing. It produces a tendency to influence the behavior of the system. Frijda calls this *control precedence*.
4. *Action readiness change*: According to Frijda, this is at the heart of the emotional response. Change in willingness to act means changes in the allocation of processing and attention resources, as well as the tendency to perform certain types of actions.
5. Regulation: In addition to activating certain forms of willingness to act, the emotional system also observes all processes of the overall system and events in the environment that can influence this willingness to act in order to be able to intervene accordingly.
6. Social nature *of the environment*: The emotional system is set up to operate in a predominantly social environment. Many assessment categories are therefore of a social nature; Willingness to act is predominantly a willingness to take social action.

For Frijda, emotions are imperative for systems that realize multiple concerns in an uncertain environment. When a situation arises in which the realisation of these concerns appears to be at risk, so-called *action tendencies arise.* These tendencies to act are closely linked to emotional states and serve to ensure the enforcement of concerns, what Frijda calls concern *realization* (CR).

Frijda (1986) defines the following as essential tendencies of action (related emotions in brackets):

- *Approach (Desire)*
- *Avoidance (Fear)*
- *Being-with (Enjoyment, Confidence)*
- *Attending (Interest)*
- *Rejecting (Disgust)*
- *Nonattending (Indifference)*
- *Agonistic (Attack/Threat) (Anger)*
- *Interrupting (Shock, Surprise)*
- *Dominating (Arrogance)*
- *Submitting (Humility, Resignation)*

In detail, according to Frijda, a functioning emotional system must have the following components:

Concerns: Internal representations against which the existing conditions are tested.

Action Repertoire: Consists of rapid emergency responses, social signals, and mechanisms to develop new plans.

Appraisal Mechanisms: Mechanisms that determine similarities between events and concerns as well as connections to the action control system and the repertoire of actions.

Analyser: Observing the incoming information and coding with regard to implications and consequences.

Comparator: Test all information for concern relevance. The result is relevance signals that activate the action system and the *diagnostician* and cause *arousal (attentional arousal*).

Diagnostician: Responsible for context evaluation, the search of information for actionable clues. Performs a series of tests (e.g., whether consequences of an event are certain or uncertain, who is responsible for them, etc.) and results in an *appraisal profile*.

Evaluator: Match or non-match signals from the *comparator* and the profile of the *diagnosher* are combined to create the final relevance signal and its intensity parameter. The intensity signals the urgency of an action to the action system. The relevance signal constitutes the so-called *control precedence signal*.

Action Proposer: Prepares the action by selecting a suitable action alternative and providing the necessary resources.

Actor: Generates actions.

This general description of an emotional system can be formalized in such a way that it can form the basis for a computer model:

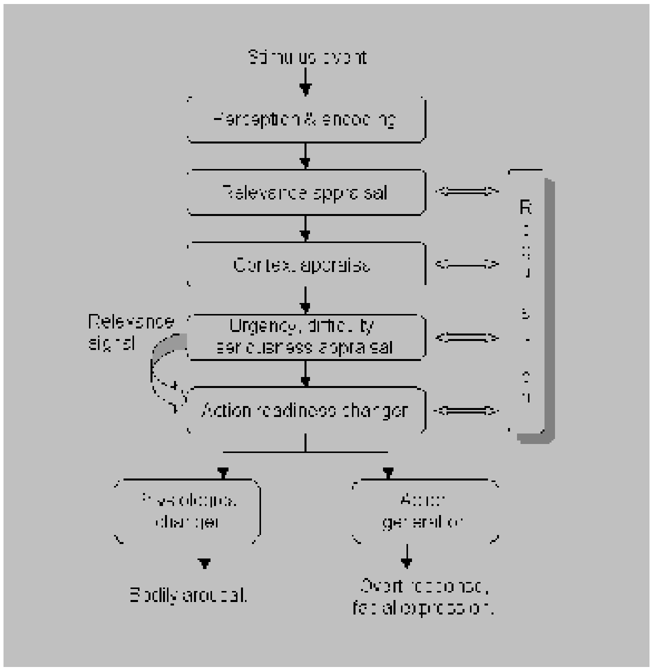


Fig. 3: Frijda's emotional system (Frijda and Moffat, 1994)

The theory outlined so far was presented by Frijda in 1986. The computer model ACRES (Frijda and Swagerman, 1987), which is described below, is also based on it. The evaluation of ACRES subsequently led Frijda to make a number of modifications to his theoretical approach. These are also presented below in connection with the computer model WILL (Moffat and Frijda, 1995).

4.5. The theory of Oatley & Johnson-Laird

Oatley and Johnson-Laird explicitly developed their theory in a form that can be implemented as a computer model, even if they did not take this step themselves. For them, the necessity for this is that almost all computer models of the human mind have not taken emotions into account, while they see them as a central component for the organization of cognitive processing processes.

In what they themselves call the "communicative theory of emotions" *(Oatley & Jenkins, 1996, p. 254), Oatley and Johnson-Laird assume* a hierarchy of parallel processing instances that work asynchronously on different tasks. These instances are coordinated by a central control system (or operating system). The control system contains a model of the entire system.

The individual modules of the system communicate with each other so that they can function at all. According to Oatley and Johnson-Laird, there are two types of communication. The first is what they call *propositional* or symbolic; they transmit factual information about the environment. The second type of communication is *nonpropositional* or emotional in nature. Their task is not to transmit information, but to put the entire system of modules in a state of increased attention, a so-called *emotion mode*. This function is comparable to that of *global interrupt* programs on computers:

> "Emotion signals provide a specific communication system which can invoke the actions of some processors [modules] and switch others off. It sets the whole system into an organized emotion mode without propositional data having to be evaluated by a high-level conscious operating system... The emotion signal simply propagates globally through the system to set into one of a small number of emotion modes."
> (Oatley & Johnson-Laird, 1987, S. 33)

According to Oatley, the central postulate of the theory is:

> "Each goal and plan has a monitoring mechanism that evaluates events relevant to it. When a substantial change of probability occurs of achieving an important goal or subgoal, the monitoring mechanism broadcasts to the whole cognitive system a signal that can set it into readiness to respond to this change. Humans experience these signals and the states of readiness they induce as emotions."
> (Oatley, 1992, p. 50)

Emotions coordinate quasi-autonomic processes in the nervous system by communicating significant roadmaps (*plan junctures*) of current plans.

Oatley and Johnson-Laird associate such milestones of plans with elemental emotions:

Plan juncture	*Emotion*
Subgoals being achieved	Happiness
Failure of major plan	Sadness
Self-preservation goal violated	Anxiety
Active plan frustrated	Anger
Gustatory goal violated	Disgust

Table 2: Milestones of plans (according to Oatley, 1992, p. 55)

Since they occur at milestones, emotions are a design solution to problems of plan change in systems with multiple goals.

The term *communicative theory of emotion* comes from the fact that it is the task of emotions to transmit very specific information to all modules of the overall system.

Following a suggestion by Sloman, Oatley has once again specified that there are two types of signals in the model: semantic signals and control signals. The two can, but do not have to, perform together. For example, Oatley (1992) claims that his model is the only one that can explain a vague emotional state: in this case, only the control signals are active, but not the semantic ones.

5. Electronic Assistants

There are a variety of models in which computers have been used to recognize or represent emotions. These models are not "emotional computers" in the true sense of the word, as their "emotional" components are predefined elements and not a subsystem that has evolved on its own.

The models described below are largely rule-based production systems. Thus, they are also symbol-processing systems at the same time. From the 1960s to the present day, there has been a lively discussion as to whether or to what extent the human mind is a symbol-processing system and to what extent symbol-processing computer models can be a realistic approximation of real-world functions (see e.g. Franklin, 1995).

A rules-based production system has a number of standard components as a minimum:

1) a so-called *knowledge base*, which contains the processing rules of the system;
2) a so-called *global database*, which represents the working memory of the system;
3) a so-called *control structure*, which analyzes the content of the *global database* and decides which processing rules of the *knowledge base* are to be applied.

Franklin (1995) provides a more detailed description of a rule-based production system using the example of SOAR: The system operates within a defined *problem space*; the production process of the system is the application of corresponding *condition-action rules* that transform the *problem space* from one state to another.

The following models by Dyer, Pfeifer, Bates and Reilly as well as Elliott can be regarded as rule-based production systems. Scherer's model breaks out somewhat in that it is an implementation that is not rule-based. However, its underlying approach is an assessment theory and could easily be implemented as a production system.

5.1. Dyer's Models

Dyer has developed a total of three models over the years: BORIS, OpEd, and DAYDREAMER. BORIS and OpEd are systems that can infer emotions from texts; DAYDREAMER is a computer model that can generate emotions.

Dyer considers emotions as an emergent phenomenon:

> “Neither BORIS, OpEd, nor DAYDREAMER were designed to address specifically the problem of emotion. Rather, emotion comprehension and emotional reactions in these models arise through the interaction of general cognitive processes of retrieval, planning and reasoning over memory episodes, goals, and beliefs.”
> (Dyer, 1987, p. 324)

In Dyer's work, these "general cognitive processes" are realized in the form of *demons*, specialized program subroutines that are activated under certain conditions and perform specific tasks independently of each other. After completing their work, these *demons* "die" or create new subroutines.

> “In BORIS, “disappointed” caused several demons to be spawned. One demon used syntactic knowledge to work out which character x was feeling the disappointment.

> Another demon looked to see if x had suffered a recent goal failure and if this was unexpected."
> (Dyer, 1987, p. 332)

5.1.1. BORIS

BORIS is based on a so-called *affect lexicon*, which has six components: a person who feels the emotion; the polarity of the emotion (positive - negative); one or more target achievement situations; the object or person at whom the emotion is directed; the strength of the emotion and the respective expectations.

With these components, emotions are represented in BORIS as follows:

Emotion:	Facilitation
Person:	x
Polarity:	positive
directed at:	-/-
Achievement of objectives:	Goal achieved
Expectation:	Expectation not fulfilled

In this case, person x did not expect to achieve his goal. This expectation has not been fulfilled. Now person x experiences a positive state of arousal, which he perceives as relief (according to Dyer, 1987, p. 325).

In a similar form, emotions such as *happy*, *sad*, *grateful*, *angry-at*, *hopeful*, *fearful*, *disappointed*, *guilty,* etc. are represented in BORIS.

This underlines the claim that Dyer is pursuing with BORIS: All emotions can be represented in BORIS in the form of a negative or positive state of arousal, combined with information about a person's goals and expectations.

Dyer points out that the variables he mentions can also be used to represent emotions for which there is no corresponding word in a particular language.

With the help of this model, BORIS can draw conclusions about the respective goal achievement situation of a person; Understand and generate text that contains descriptions of emotions, as well as tap into and compare the meanings of emotional terms. The system is also able to represent multiple emotional states.

From the goal/plan analysis carried out by a person and the resulting result, BORIS can also develop expectations of how this person will continue to behave in order to achieve his or her goals. The strength of a state of excitement can also be used by BORIS for such predictions.

5.1.2. OpEd

OpEd is an extension of BORIS. While BORIS can only understand emotions in narrative texts due to its internal lexicon, OpEd is able to infer emotions and beliefs from non-narrative texts as well:

> “OpEd is... designed to read and answer questions about editorial text. OpEd explicitly tracks the beliefs of the editorial writer and builds representations of the beliefs of the writer and of those beliefs the writer ascribes to his opponents.”
> (Dyer, 1987, p. 329)

Beliefs are implemented in OpEd based on four dimensions: *Believer* is the one who has a certain belief; *content* is an evaluation of goals and plans; *attack* are the beliefs that oppose the currently expressed one; *support* are the beliefs that support the current belief.

According to Dyer, *beliefs were* an essential element that was missing from BORIS. For example, the statement "happy(x)" in BORIS is represented as the achievement of a goal by x. According to Dyer, this is not enough:

> “What should have been represented is that happy(x) implies that x *believes* that x has achieved (or will achieve, or has a chance of achieving) a goal of x.”
> (Dyer, 1987, p. 330)

That's why OpEd adds new demons to the demons known from BORIS : *belief-building, affect-related demons*.

Dyer has shown that OpEd is able not only to infer the author's beliefs from newspaper texts, but also to draw conclusions about the beliefs of those against whom the author takes a stand.

5.1.3. DAYDREAMER

While BORIS and OpEd are designed to understand emotions, DAYDREAMER (Mueller and Dyer, 1985) is an attempt to develop a system that "feels" emotions. This sensation does not manifest itself in a subjective state of the system, but in the fact that its respective "emotional" state influences its internal behavior in the processing of information.

Mueller and Dyer define four essential functions of daydreaming: they increase the effectiveness of future behavior by anticipating possible reactions to expected events; they support learning from successes and mistakes by playing through alternative courses of action; they support creativity because the imaginary replaying of action sequences can lead to new solutions, and they support the regulation of emotions by reducing their perceived intensity.

To achieve these goals, DAYDREAMER is equipped with the following main components:

1. a *scenario generator*, which *consists of* a planner *and so-called* relaxation rules;
2. a dynamic episodic memory, the contents of which are used by the *scenario generator* ;
3. a collection of personal goals and *control goals* that guide the *scenario generator*;
4. an emotion component, in which daydreams arise or are initiated by emotional states triggered by the achievement or failure to achieve goals;
5. domain *knowledge* of interpersonal relationships and everyday activities.

DAYDREAMER has two types of functions, the *daydreaming mode* and the *performance mode*. In *daydreaming mode* , the system moves continuously in daydreams until it is interrupted; in *performance mode* , the system shows what it has learned from daydreaming.

Mueller and Dyer postulate a set of goals that a system has, which they call control goals. These are partly triggered by emotions and in turn trigger daydreams. The function of control goals is to provide a short-term modification of emotional states and to ensure the achievement of personal goals in the long term.

The system thus has a feedback mechanism in which emotions trigger daydreams and daydreams modify these emotions and trigger new emotions, which in turn initiate new daydreams.

Mueller and Dyer name four control goals that occur when daydreaming:

a. *Rationalization:* The goal of rationalizing away a failure and thus reducing a negative emotional state.
b. *Revenge:* The goal of preventing someone else from achieving a goal and thus reducing one's own anger.
c. *Reversal of success or failure:* The goal of imagining a scenario with an opposite outcome in order to reverse the polarity of an emotional state.
d. *Preparation:* The goal of developing hypothetical episodes to play out the consequences of a possible plot.

Mueller and Dyer describe the functioning of DAYDREAMER with an example in which DAYDREAMER depicts an active young man with social goals who has met an actress who has declined his invitation for a drink.

DAYDREAMER then generates the following two daydreams:

> "*Daydream 1*: I am disappointed that she didn't accept my offer... I imagine that she accepted my offer and we soon become a pair. I help her when she has to rehearse her lines... When she has to do a film in France, I drop my work and travel there with her... I begin to miss my work. I become unhappy and feel unfulfilled. She loses interest in me, because I have nothing to offer her. It's good I didn't get involved with her, because it would've led to disaster. I feel less disappointed that she didn't accept my offer.
> (......)
> *Daydream 2*: I'm angry that she didn't accept my offer to go have a drink. I imagine I pursue an acting career and become a star even more famous than she is. She remembers meeting me a long time ago in a movie theater and calls me up... I go out with her, but now she has to compete with many other women for my attention. I eventually dump her."
> (Dyer, 1987, p. 337)

The first daydream is an example of reversal: it pretends that the rendezvous has taken place and develops a fantasy about the consequences. The reality monitor reports that an important goal, namely one's own career, is neglected. The result is a rationalization that reduces the negative emotional state.

Daydream 2 is triggered by the emotional state of anger and embodies revenge in order to reduce the negative effect of the current emotional state.

Once a control target is activated, the *scenario generator* generates a series of events related to the control target. These daydreams differ from classic plans in that they are not doggedly focused on a goal, but can change in loose, associative sequence. The system has a relaxation mechanism for this purpose, which also enables unrealistic daydreams. Mueller and Dyer give four examples of such relaxations in their model:

- *Behavior of others*: DAYDREAMER can assume that the movie star accepts his offer.
- *Self attributes*: DAYDREAMER can assume to be a high-performance athlete or a well-known movie star.
- *Physical constraints*: You can assume that you are invisible or flying.
- *Social constraints*: You can assume that you are making a scene in a posh restaurant.

The strength of the relaxations is not always the same; it varies according to the respective active control goals.

Positive emotions occur through the memory of achieving a goal, negative emotions through the memory of failure. If someone else is responsible for not achieving a goal of DAYDREAMER, the emotion *anger* is triggered. Imaginary successes imagined in daydreams evoke positive emotions; imaginary failures, negative emotions.

During his daydreams, DAYDREAMER stores complete daydreams, future plans and planning strategies in his memory. These are indexed in episodic memory and can be retrieved later. This allows the system to learn from its daydreams for future situations.

A computer's ability to daydream is essential to the development of its intelligence, Mueller and Dyer claim. They therefore imagine computers that can daydream during the time they are not being used in order to increase their performance in this way.

The model of Mueller and Dyer, as far as it can be judged here, has not been further developed according to its original idea

5.2. The Pfeifer Model

Pfeifer (1982, 1988) presented FEELER ("Framework for Evaluation of Events and Linkages into Emotional Responses"), a model of an emotional computer system that explicitly refers to emotional psychological theoretical approaches.

Pfeifer's model is a rule-based system with *working memory (WM), long-term memory (LTM*) and control structure; however, he also distinguishes between declarative and procedural knowledge *when it comes to the contents of long-term memory (the* knowledge base).

In order to be able to represent emotions, Pfeifer expands this structure of a rule-based system with further subsystems. Thus, FEELER not only has a cognitive, but also a physiological working memory.

In order for emotions to arise in FEELER, the system needs a schema to analyze the cognitive conditions that lead to an emotion. For this purpose, Pfeifer uses the taxonomy developed by Weiner (1982). From this, he develops an exemplary rule for the emergence of an emotion:

```
"IF    current_state is negative for self
       and emotional_target is VARperson
       and locus_of_causality is VARperson
       and locus_of_control is VARperson
```

```
THEN  ANGER at VARperson”
```

(Pfeifer, 1988, S. 292)

In order for this rule to take effect, all its requirements must first be represented in the World Cup. This is done through inference processes that store their results in the WM. According to Pfeifer, such inference processes are typically triggered by interrupts.

Corresponding interrupts are generated in FEELER when expectations with regard to the achievement of subgoals are violated or when there are no expectations for an event.

In a second rule, Pfeifer defines a tendency to act that follows Rule 1:

```
IF    Angry
      and emotional_target is VARperson
      and int_pers_rel self - VARperson is negative
THEN  generate goal to harm VARperson
```

(Pfeifer, 1988, S. 297)

According to Pfeifer, this rule also makes the heuristic value of an emotion clear: the emotion reduces the circle of possible candidates and actions for inference processes.

Pfeifer himself admits that such a model is not able to cover all emotional states. He discusses a number of problems, for example the interaction of different subsystems and their influence on the arising, duration and decay of emotions. In a further step, Pfeifer supplemented his model with the taxonomy of Roseman (1979) in order to be able to represent emotions in FEELER in connection with the achievement of goals.

5.3. The Bates and Reilly Model

In his essay "The Role of Emotion in Believable Agents" (Bates, 1994), Joseph Bates quotes Disney cartoonist Chuck Jones as saying that Disney always strives for believability in its cartoon characters. Bates continues:

> “Emotion is one of the primary means to achieve this believability, this illusion of life, because it helps us know that characters really care about what happens in the world, that they truly have desires.”
> (Bates, 1994, S. 6)

Together with a number of colleagues, Bates has launched the Oz Project at Carnegie-Mellon University . The goal is to produce synthetic creatures that should appear as lifelike as possible to their human audience. In short, it is about an interactive drama system or "artistically effective simulated worlds" (Bates et. al., 1992, p.1).

The basic approach is to create *broad and shallow agents*. While computer models of AI and emotions focus on specific areas and try to cover them as intensively as possible, Bates takes the opposite approach:

> “... part of our effort is aimed at producing agents with a broad set of capabilities, including goal-directed reactive behavior, emotional state and behavior, social knowledge and behavior, and some natural language abilities. For our purpose, each of these capacities can be as limited as is necessary to allow us to build broad, integrated agents...”
> (Bates et. al., 1992a, S.1)

According to Bates, the broad approach is necessary to create believable artificial characters. Only an agent who is able to respond convincingly to a variety of situations in an environment that includes a human user will be truly accepted as a credible character by the latter.

Since *Oz* is deliberately designed as an artificial world that should be viewed by the user as a film or play, it is sufficient to lay out the system's multiple capabilities "flat" to satisfy the user's expectations. Because as in the cinema, it does not expect a correct depiction of reality, but an artificial world with credible actors in this context.

An *Oz* world consists of four essential elements: a simulated environment, a number of agents who populate this artificial world, an interface through which people can participate in what is happening in the world, and a planner who deals with the long-term structure of a user's experiences.

Bates' agent architecture is called *Tok* and consists of a number of components: There are modules for goals and behavior, for sensory perception, speech analysis and speech generation. And there is a module called *Em* for emotions and social relationships.

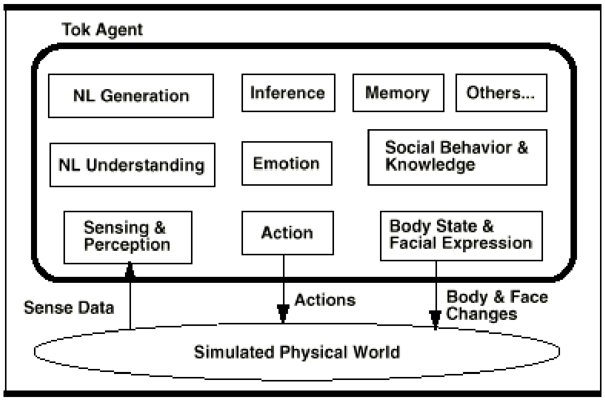


Fig. 4: Structure of the TOK architecture (Reilly, 1996, p. 14)

Em has an emotion system based on the model of Ortony, Clore and Collins (1988). However, the OCC model is not implemented in Em in all its complexity . This applies in particular to the intensity variables postulated by Ortony, Clore and Collins and their complex interaction. *Em* uses a simpler subset of these variables that is considered sufficient for the intended purpose.

Reilly (1996) explains that the use of such subsets does not reduce the OCC model, but extends it. He illustrates this with two examples:

For Ortony, Clore, and Collins, pity is generated as follows: Agent A feels sorry for Agent B when Agent A likes Agent B and Agent A considers an event to be unpleasant for Agent B in terms of its goals. "So, if Alice hears that Bill got a demotion, Alice must be able to match this event with a model of Bill's goals, including goals about demotions." (Reilly, 1996, p. 53) This would mean that Alice would have to have a relatively comprehensive knowledge of Bill's goals and assessment mechanisms - a difficult task, according to Reilly, in a dynamic world in which goals can change quickly.

Instead, he suggests the following mechanism: Agent A feels sorry for Agent B when Agent A likes Agent B and Agent A believes Agent B is unhappy. According to Reilly, this description not only has the advantage of being simpler:

> “In this case, I have broken the OCC model into two components: recognizing sadness in others and having a sympathetic emotional response..... Recognizing sadness in others is done, according to the OCC model, only through reasoning and modeling of the goals of other agents, so this inference can be built into the model of how the emotion is generated. Em keeps the recognition of sadness apart from the emotional response, which allows for multiple ways of coming to know about the emotions of others. One way is to do reasoning and modeling, but another way, for example, is to see that an agent is crying.
> The Em model is more complete than the OCC model in cases such as agent A seeing that agent B is sad but not knowing why. In the OCC case, when agent A does not know why agent B is unhappy, the criteria for pity is not met. Because the default Em emotions generators require only that agent A believe that agent B is unhappy, which can be perceived in this case, Em generates pity.”
> (Reilly, 1996, S. 53f.)

As a second example, Reilly (1996) cites the development of *distress*. In the OCC model, *distress* occurs when an event is considered unpleasant in terms of an agent's goals. This means that external events must be evaluated. In *Em,* distress *is* caused by the fact that goals are either not achieved or the probability that they will not be achieved increases, which is related to the motivation and action system. Reilly elaborates:

> “This shifts the emphasis towards the goal processing of the agent and away from the cognitive appraisal of external events. This is useful for two reasons. First, the motivation system is already doing much of the processing (e.g., determining goal successes and failures), so doing it in the emotion system as well is redundant. Second, much of this processing is easier to do in the motivation system since that’s where the relevant information is. For instance, deciding how likely a goal is to fail might depend on how far the behavior to achieve that goal has progressed or how many alternate ways to achieve the goal are available - this information is already in the motivation system.”
> (Reilly, 1996, S. 54f.)

In this way, emotional structures are to be created that are more complete and easier to apply than their purely cognitive role models. The following table shows which emotions can be generated by Em and on what basis:

Emotion Type	**Cause in Default Em System**
Distress	Goal fails or becomes more likely to fail and it is important to the agent that the goal not fail.
Joy	Goal succeeds or becomes more likely to succeed and it is important to the agent that the goal succeed.
Fear	Agent believes a goal is likely to fail and it is important to the agent that the goal not fail.
Hope	Agent believes a goal is likely to succeed and it is important to the agent that the goal succeed.
Satisfaction	A goal succeeds that the agent hoped would succeed.
Fears-Confirmed	A goal failed that the agent feared would fail.
Disappointment	A goal failed that the agent hoped would succeed.
Relief	A goal succeeds that the agent feared would fail.
Happy-For	A liked other agent is happy.
Pity	A liked other agent is sad.
Gloating	A disliked other agent is sad.
Resentment	A disliked other agent is happy.
Like	Agent is near or thinking about a liked object or agent.
Dislike	Agent is near or thinking about a disliked object or agent.
Other attitude-based emotions	Agent is near or thinking about an object or agent that the agent has an attitude towards (e.g., awe).
Pride	Agent performs an action that meets a standard of behavior.

Shame	Agent performs an action that breaks a standard of behavior.
Admiration	Another agent performs an action that meets a standard of behavior.
Reproach	Another agent performs an action that breaks a standard of behavior.
Anger	Another agent is responsible for a goal failing or becoming more likely to fail and it is important that the goal not fail.
Remorse	An agent is responsible for one of his own goals failing or becoming more likely to fail and it is important to the agent that the goal not fail.
Gratitude	Another agent is responsible for a goal succeeding or becoming more likely to succeed and it is important that the goal succeed.
Gratification	An agent is responsible for one of his own goals succeeding or becoming more likely to succeed and it is important to the agent that the goal succeed.
Frustration	A plan or behavior of the agent fails.
Startle	A loud noise is heard.

Table 3: Types of emotions and their emergence in *Em* (according to Reilly, 1996, p. 58 f.)

Reilly explicitly points out that these types of emotions do not claim to be psychologically correct, but merely represent a starting point for creating believable emotional agents.

The emotion types of *Em* are arranged in the following hierarchy:

	Positive	Joy Hope Happy-For Gloating Love Satisfaction Relief Pride Admiration Gratitude Gratification	
Total			
	Negative	Distress Fear Pity Resentment Hate Disappointment Fears-Confirmed Shame Reproach Anger Remorse	Startle Frustration

Table 4: Hierarchy of Emotion Types in *Em* (according to Reilly, 1996, p. 76)

It is noticeable that in this hierarchy the emotion types based on the OCC model are arranged one level below the level "positive - negative". This *mood level* gives *Em* the opportunity to determine the general mood of an agent before a more in-depth analysis, which greatly simplifies the production of emotional effects.

To determine the general mood (*good-mood vs. bad-mood),* Em first *sums* up the intensities of the positive emotions, then those of the negative emotions. Formalized, it looks like this:

$$I_p = \log_2 (\sum_e 2^{I_e}),\ e \in \{positive\,emotions\}$$
$$I_n = \log_2 (\sum_e 2^{I_e}),\ e \in \{negative\,emotions\}$$

```
IF      $Ip > In$
THEN    set $good\text{-}mood = Ip$
AND     set $bad\text{-}mood = 0$
ELSE    set $good\text{-}mood = 0$
AND     set $bad\text{-}mood = - In$
```

(according to Picard, 1997, p. 202)
The TOK system has been implemented with different characters. One of the best known is Lyotard, a virtual domestic cat. Bates et al. (1992b) describe a typical interaction with Lyotard:

> “As the trace begins, Lyotard is engaged in exploration behavior in an attempt to satisfy a goal to amuse himself... This behavior leads Lyotard to look around the room, jump on a potted plant, nibble the plant, etc. After suffcient exploration, Lyotard's goal is satisfied. This success is passed on to Em which makes Lyotard

> mildly happy. The happy emotion leads to the “content" feature being set. Hap then notices this feature being active and decides to pursue a behavior to find a comfortable place to sit, again to satisfy the high-level amusement goal. This behavior consists of going to a bedroom, jumping onto a chair, sitting down, and licking himself for a while.
> At this point, a human user whom Lyotard dislikes walks into the room. The dislike attitude, part of the human-cat social relationship in Em, gives rise to an emotion of mild hate toward the user. Further, Em notices that some of Lyotard's goals, such as not-being-hurt, are threatened by the disliked user's proximity. This prospect of a goal failure generates fear in Lyotard. The fear and hate combine to generate a strong “aggressive" feature and diminish the previous “content" feature.
> In this case, Hap also has access to the fear emotion itself to determine why Lyotard is feeling aggressive. All this combines in Hap to give rise to an avoid-harm goal and its subsidiary escape/run-away behavior that leads Lyotard to jump off the chair and run out of the room.”
> (Bates et al., 1992b, S. 7)

Reilly (1996) tested the credibility of a virtual character equipped with Em. Test subjects were confronted with two virtual worlds in which two virtual characters acted. The difference between the two worlds was that in one case both characters were equipped with *Em*, while in the second case only one character had it.

A questionnaire was then used to determine the differences perceived by the test subjects between the *em* character ("Melvin") and the non-Em character ("Chuckie").

The test subjects rated Melvin as more emotional than Chuckie. His credibility was also rated higher than the Chuckies. At the same time, the test subjects stated that Melvin's personality was more contoured than that of Chuckie and that they felt less often in Melvin than in Chuckie that they were dealing with fictional characters.

However, the significance of the results varies considerably, so that Reilly also admits that *Em* is only "moderately successful" (Reilly, 1996, p. 129).

5.4. The Model of Elliott

Another model based on the theory of Ortony, Clore and Collins is Clark Elliott's *Affective Reasoner*. Elliott is primarily interested in the role of emotions in social interactions, whether between people, between humans and computers, or between virtual actors in a virtual computer world.

Elliott summarizes the core elements of the *Affective Reasoner* as follows:

> “One way to explore emotion reasoning is by simulating a world and populating it with agents capable of participating in emotional episodes. This is the approach we have taken. For this to be useful we must have (1) a simulated world which is rich enough to test the many subtle variations a treatment of emotion reasoning requires, (2) agents capable of (a) a wide range of affective states, (b) an interesting array of interpretations of situations leading to those states and (c) a reasonable set of reactions to those states, (3) a way to capture a theory of emotions, and (4) a way for agents to interact and to reason about the affective states of one another. The Affective Reasoner supports these requirements.”
> (Elliott, 1992, S. 2)

According to Elliott, the advantages of such a model are manifold: On the one hand, it makes it possible to test psychological theories about the origin of emotions and the resulting actions for their internal plausibility and stringency. Secondly, affective modules are an important component of distributed agent systems if they are to act in real time without friction losses. Third, a computer model that can understand and express emotions is an essential step in designing better human-machine interfaces.

As an example of a simulated world, Elliott (1992) chooses *Taxiworld*, a scenario with four taxi drivers in Chicago. *(Taxiworld* is not limited to four drivers; the simulation was carried out with up to 40 drivers.) There are different stops, different passengers, police officers and different destinations. This can be used to create a series of situations that lead to the emergence of emotions.

Taxi drivers must be able to interpret these situations in such a way that emotions can arise. To do this, they need the ability to reflect on the emotions of other taxi drivers. After all, drivers should also be able to act on the basis of their emotions.

Elliott illustrates the difference between the *affective reasoner* and classical analysis models of AI with the following example (Elliott, 1992): "Tom's car didn't start, and Tom missed an appointment because of it. He insulted his car. Harry observed this incident."

A classic AI system would draw the following conclusions from this story: Tom should have his car repaired. Harry has learned that Tom's car is defective. Tom couldn't get to his appointment on time without his car. Harry suggests that Tom leave early for his dates in the future.

The *Affective Reasoner* , on the other hand, would come to completely different conclusions: Tom blames his car for his missed appointment. Tom is angry. Harry can't understand why Tom is angry with his car, since you can't blame a car for anything. Harry advises Tom to calm down again. Harry feels sorry for his friend Tom because he is so aroused.

In order to be able to react in this way, the *affective reasoner* needs a relatively large number of components. Although he specializes in emotions, Elliott still refers to him as a "*shallow model*" (Elliott, 1994). In the following, the essential components of the *affective reasoner* as described by Elliott (1992) will be presented.

5.4.1. The Construction of Agents

The agents of the *Affective Reasoner* have a rudimentary personality. This personality consists of two components: the *interpretive personality component* represents the individual disposition of an agent to interpret situations in his world. The *manifestative personality component* is his individual way of showing his emotions.

Each agent has one or more *goals*. This refers to situations that the agent considers desirable to occur. In order to be able to act emotionally, the agents need an *object domain* in which situations occur that lead to emotions and in which the agents can perform actions triggered by emotions.

Every agent needs several databases to function and must have access to them at all times:

1. A database of 24 emotion types, which essentially correspond to the emotion types of Ortony, Clore and Collins (1988) and was expanded by Elliott to include the two types love and *hate*. Each of these emotion types is assigned special *emotion eliciting conditions* (ECC).

2. A database of *goals*, standards and *preferences*. These GSPs constitute the *concern structure* of an agent and at the same time define its *interpretive personality component*.

3. A database of accepted GSPs for other agents in his world. Elliott calls it a COO (*Concerns-of-Others)* database. Since this is data learned by the agent, it is usually imperfect and can also contain incorrect assumptions.

4. A database of reaction patterns, which are divided into up to twenty different groups, depending on the type of emotion.

5.4.2. The generation of emotions

The patterns stored in the GSP and COO databases are compared by the agent to the EECs in their world, and a group of bonds is created in correspondence. Some of these bonds represent two or more values for a class that Elliott calls *emotion eliciting condition relation* (*EEC relation*).

EEC relations are composed of elements of the situation that evokes the emotion and its interpretation by the agent. Taken together, this can create the prerequisite for the invocation of an emotion:

self	other	desire-self	desire-other	pleasingness	status	Evaluatestion	Respondssible agent	appealingness
(*)	(*)	(d/u)	(d/t)	(p/d)	(u/c/d)	(p/b)	(*)	(a/u)

Key to attribute values	
abbreviation	**meaning**
*	some agent's name
d/u	desirable or undesirable (event)
p/d	pleased or displeased about another's fortunes (event)
p/b	praiseworthy or blameworthy (act)
a/u	appealing or unappealing (object)
u/c/d	unconfirmed, confirmed or disconfirmed

Tab. 5: *EEC relations* des *Affective Reasoner* (nach Elliott, 1992, S. 37)

If one or more *EEC relations* are formed, they are used to generate emotions. In this phase, a number of problems arise, which are discussed in detail by Elliott because they have not been sufficiently taken into account in the theory of Ortony, Clore and Collins.

As an example, Elliott cites a compound emotion. The *affective reasoner* constructs the *EEC relations* for the two underlying emotions and then summarizes them in a new *EEC relation*. The constituent emotions are thus replaced by the compound emotion. Elliott doesn't see this as the optimal solution:

> "Does anger really subsume distress? Do compound emotions always subsume their constituent emotions? That is, in feeling anger does a person also feel distress and reproach? This is a diffcult question. Unfortunately, since we are implementing a platform that generates discrete instances of emotions, we cannot finesse this issue. Either they do or they do not. There can be no middle ground until the eliciting condition theory is extended, and the EEC relations extended."

(Elliott, 1992, S. 42)

While this approach may still work for qualitatively similar emotions (Elliott cites *distress* and *anger* as examples), a problem arises at the latest when several emotions occur at the same time, especially if they contradict each other.

With several instances of the same emotion, the solution is still quite simple. For example, if an agent has two goals in the card game ("win" and "earn money"), his win in the card game triggers the emotion *happy* twice . The *affective reasoner* then simply generates two instances of the same emotion.

The situation is more problematic with contradictory emotions. Elliott acknowledges that there are gaps in the OCC model at this point, stating, "Except for the superficial treatment of conflicting expressions of emotions, the development and implementation of a theory of the expression of multiple emotions is beyond the scope of this work." (Elliott, 1992, p. 44f.) The *affective reasoner* therefore shifts the "solution" of this problem to its action-generation module (see below).

5.4.3. The generation of actions

Once an emotional state has been generated for an agent, a resulting action is initiated. The *affective reasoner* uses an *emotion manifestation lexicon* that has three dimensions: the 24 emotion types, the twenty or so reaction types (*emotion manifestation categories*) and an intensity hierarchy of possible reactions (which was not implemented in the first model of the *affective reasoner*).

The reaction types of the *affective reasoner* are based on a list by Gilboa and Ortony (unpublished). These are hierarchically organized; each hierarchical level is further structured along a continuum from spontaneous to planned reactions. As an example, Elliott cites the action categories for *"gloating* ":

Spontaneous	Non goal-directed	Expressive	Somatic	flush, tremble, quiet pleasure
			Behavioral (towards inanimate)	slap
			Behavioral (towards animate)	smile, grin, laugh
			Communicative (non verbal)	superior smile, throw arms up in air
			Communicative (verbal)	crow, inform-victim
		Information Processing	Evaluative self-directed attributions of...	superiority, intelligence, prowess, invincibility
			Evaluative agent-directed attributions of....	silliness, vulnerability, inferiority
			Obsessive Attentional focus on...	other agent's blocked goal

Gloating	Goal directed	Affect-oriented Emotion regulation and modulation	Repression Reciprocal Suppression Distraction Reappraisal of self as.... Reappraisal of situation as... Other-directed emotion modulation	deny positive valence “rub-it-in” show compassion focus on other events winner modifiable, insignificant induce embarrassment, induce fear, induce sympathy for future, induce others to experience joy at victim’s expense
Planned		Plan-oriented	Situated plan-initiation Full plan-initiation	call attention to the event plan for recurrence of event

Table 6: Reaction types of the *affective reasoner* for *"gloating"* (according to Elliott, 1992, p. 97)

For each agent, individual categories can be activated or deactivated before the simulation begins. This specific pattern of active and non-active categories constitutes the individual *manifestative personality* of an agent. Elliott refers to the activated categories as the potential *temperament traits* of an agent.

In order to avoid conflicts between contradictory emotions and thus also contradictory actions, the action module contains so-called *action exclusion sets*. They are formed by dividing the possible reactions into equivalence classes. A member of one of these classes can never appear together with a member of another class in the final *action set* .

5.4.4. Interpreting the emotions of other agents

An agent obtains his knowledge of the emotions of other agents not only through pre-programmed characteristics, but also by observing other agents within the simulation and drawing his conclusions from them. These then flow into his COO database. To integrate this learning process into the *Affective Reasoner* , Elliott uses a program called *Protos* (Bareiss, 1989).

One agent observes the emotional reaction of another agent. *Protos* then allows the agent to draw conclusions about the emotion that the other agent feels and thus demonstrate empathy.

First of all, the observed emotional response is compared to a database of emotional reactions to define the underlying emotion. Then, the observed event is filtered through the COO database for the observed agent to determine whether this response has already been recorded. If this is the case, it can be assumed that the database contains a correct representation of the emotion-triggering situation. Based on this, the observing agent can then develop an explanation for the observed agent's behavior.

If the representation in the COO database does not match the observed behavior, it is removed from the database and the database is searched again. If no correct representation is found, the agent can fall back on default values, which are then integrated into the COO database.

Since COOs are nothing more than assumed GSPs for another agent, the affective reasoner is able *to represent beliefs of one agent over the assumptions of another agent with the* help of so-called satellite COOs.

5.4.5. The evolution of the model

The model described in its basic features so far was presented in this form by Elliott in his dissertation in 1992. In the years that followed, he developed the *Affective Reasoner* in a number of areas.

For example, the original model lacked a component that determines the intensity of emotions. In another work, Elliott (Elliott and Siegle, 1993) develops a group of *emotion intensity variables* based on the work of Ortony, Clore and Collins and Frijda.

Elliott divides the intensity variables into three categories. Each variable is assigned limits within which it can move (sometimes bipolar). Most intensities can take a value between 0 and 10. Weaker *modifiers* can take values between 0 and 3, *modifiers* that only reduce an intensity can only take values between 0 and 1. Variables whose effects on the intensity calculations are determined by the valence of an emotion (for example, a variable that increases the intensity of a negatively valenced emotion but decreases the intensity of a positively valenced emotion) can take values between 1 and 3 and also receive a *bias value* that determines the direction. The intensity variables and their value leeway are listed below:

a. *simulation-event variables* are variables whose values change independently of the agents' interpretation mechanisms (*goal realization/blockage*: -10 to +10, *blameworthiness-praiseworthiness*: -10 to +10, *appealingness*: 0 to 10, *repulsiveness*: -10 to 0, *certainty*: 0 to 1, *sense-of-reality*: 0 to 1, *temporal proximity*: 0 to 1, *surprisingness*: 1 to 3, effort: 0 to 3, *deservingness*: 1 to 3);
b. *stable disposition variables* haben zu tun mit der Interpretation einer Situation durch einen Agenten, sind relativ konstant und konstituieren die Persönlichkeit eines Agenten (*importance to agent of achieving goal*: 0 bis 10, *importance to agent of not having goal blocked*: 0 bis 10, *importance to agent of having standard upheld*: 0 bis 10, *importance to agent of not having standard violated*: 0 bis 10, *influence of preference on agent*: 0 bis 10, *friendship-animosity*: 0 bis 3, *emotional interrelatedness of agents*: 0 bis 3);
c. *mood-relevant variables* are fleeting, change the interpretation of a situation for an agent, can be the result of previous affective experiences and return to their default values after a certain time (*arousal*: 0.1 to 3, *physical well-being*: 0.1 to 3, *valence bias*: 1 to 3, *depression-ecstasy*: 1 to 3, *anxiety-invincibility*: 1 to 3, *importance of all Goals, Standards, and Preferences*: 0.3 to 3, *liability-creditableness*: 1 to 3).

Elliott (Elliott and Siegle, 1993) reports that an analysis of emotional episodes with the help of these variables led to the result that all emotions can be represented and recognized within the framework of the model.

In the further course, Elliott (Elliott and Carlino, 1994) extended the *Affective Reasoner* with a speech recognition module. The system was presented with sentences with emotion words, intensity modifiers and pronomial references to third parties ("*I am a bit sad because he...*"). Of 198 emotion words, 188 were already recognized in the first run. In another experiment, the sentence

"Hello Sam, I want to talk to you" was presented to the system in seven emotionally different accents (*anger, hatred, sadness, love, joy, fear, neutral*). After some training, the *Affective Reasoner* came to a 100 percent correct identification of the underlying emotion category.

In a further step, the Affective Reasoner received a module with which it can represent emotion types as facial expressions of a cartoon face (Elliott, Yang and Nerheim-Wolfe, 1993). The presentation capabilities include the 24 emotion types in three intensity levels each, each of which can be represented by one of seven schematic faces. The faces have been integrated into a morphing module that is able to perform rudimentary lip movements and fluently switch from one facial expression to the next. In addition, the *Affective Reasoner* received a speech output module and the ability to select and play different music from an extensive database depending on the emotion.

The ability of the system to correctly represent emotions was tested by Elliott (1997a) in an experiment in which 141 test subjects participated. The test subjects were shown videos in which either an actor or the faces of the *affective reasoner* recited a sentence that could have different meanings depending on the emphasis and facial expression. The actor has been thoroughly trained beforehand to express even subtle differences between emotions; the *affective reasoner* was only given the emotion category and the text. The subjects' task was to assign the correct emotional meaning to the spoken sentence from a list of alternatives. An example:

> “For example, in one set, twelve presentations of the ambiguous sentence, “I picked up Catapia in Timbuktu," were shown to subjects. These had to be matched against twelve scenario descriptions such as, (a) Jack is proud of the Catapia he got in Timbuktu because it is quite a collector's prize; (b) Jack is gloating because his horse, Catapia, just won the Kentucky Derby and his archrival Archie could have bought Catapia himself last year in Timbuktu; and (c) Jack hopes that the Catapia stock he picked up in Timbuktu is going to be worth a fortune when the news about the oil elds hits; [etc., (d) - (l)].”
> (Elliott, 1997a, S. 3)

The test subjects also indicated on a scale of 1 to 5 how sure they were of their judgments. The computer editions were divided into three groups: facial expression, facial expression and speech and facial expression, speech and background music.

Overall, the test subjects were significantly better able to correctly identify the underlying scenarios with the computer faces than with the actor (70 percent compared to 53 percent). There were hardly any differences between the three forms of representation of the computer (face: 69 percent; Face and speech: 71 percent; face, speech and music: 70 percent).

Elliott is currently working on integrating the *Affective Reasoner* as a module into two existing interactive computer teaching systems (STEVE and Design-A-Plant) to enable virtual tutors to understand and express emotions and thus make the teaching process more effective (Elliott et al., 1997).

5.5. Scherer's Model

Scherer has implemented his theoretical approach in the form of an expert system called GENESE (Geneva Expert System on Emotions) (Scherer, 1993). The motive for this was to gain further insights for emotion psychology modeling, in particular to determine how many evaluation criteria are at least necessary to unambiguously identify an emotion:

> “As shown earlier, the question of how many and which appraisal criteria are minimally needed to explain emotion differentiation is one of the central issues in research on emotion-antecedent appraisal. It is argued here that one can work towards settling the issue by constructing, and continuously refining, an expert system that attempts to diagnose the nature of an emotional experience based exclusively on information about the results of the stimulus or event evaluation processes that have elicited the emotion.”
> (Scherer, 1993, S. 331)

The system consists of a knowledge base that records which types of assessments are related to which emotions. The different assessment dimensions are linked to 14 different emotions with the help of *weights*. These weightings represent the probability with which a certain assessment is linked to a certain emotion.

The user of the program has to answer 15 questions related to a specific emotional experience, for example: "Did the situation that caused your emotion occur very suddenly or abruptly?". The user can answer each question using a quantitative scale from 0 ("Doesn't apply") to 5 ("Extraordinary").

Once all questions have been answered, the system compares the user's response pattern with the response patterns that are theoretically associated with a particular emotion. It then presents the user with a list of all 14 emotions, ordered from "Most likely" to "Most unlikely". If the computer has correctly determined the emotion, it receives confirmation from the user; if not, the user enters "incorrect". The system then presents him with another ranking of emotions. If this is also incorrect, the user enters the correct emotion and the program uses this answer to construct a specific assessment emotion database specifically for that user.

In an empirical test of the predictive power of his system, Scherer found that it worked correctly in 77.9% of all cases. Certain emotions (e.g., despair/grief) were correctly predicted more often than others (e.g., fear/worry).

Scherer's GENESIS is unusual in that it does not represent a classical rule-based system, but works with weightings in a multidimensional space. There are exactly 15 dimensions that correspond to the 16 assessment dimensions from Scherer's emotion model. Each of the 14 emotions occupies a very specific point in this room. The program makes its predictions by converting the user's responses into a point in this vector space as well, and then measuring the distances to the points for the 14 emotions. The emotion closest to the input is then named first.

It is precisely this approach that prompted Chwelos & Oatley (1994) to criticize the system. First of all, they point out that such a space with 15 dimensions can contain a total of 4.7 x 10^{11} points. This can lead to the fact that the point calculated according to the user's inputs can be far away from each of the 14 emotions. Nevertheless, the system names the closest emotion. Chwelos & Oatley argue that in such a case the answer should be "no emotion" and suggest that the system be extended by a threshold within which a given input point must be around an emotion in order to trigger a concrete response.

Secondly, they criticize the fact that the model is based on the assumption that every emotion corresponds to exactly one point in this space. They raise the question of why this is so, since different combinations of assessment dimensions can evoke one and the same emotion.

Third, Chwelos & Oatley critically examine the heuristic adjustments of the assessment dimensions implemented in GENESE, which are not to be found in Scherer's formulated model. They suspect that this could be an artifact of the vector-space approach and note that there is no theoretical motivation for it.

Finally, Chwelos & Oatley doubt that Scherer's system actually provides information about the minimum number of dimensions of assessment necessary to clearly differentiate an emotion.

5.6. The model of Frijda and Swagerman

There are two implementations of Frijda's *concern realisation* theory in computer models: ACRES (Frijda and Swagerman, 1987) and WILL (Moffat and Frijda, 1995).

ACRES (*Artificial Concern REalisation System*) is a computer program that stores facts about emotions and works with them. Frijda and Swagerman wanted to answer the initial question they had asked themselves with ACRES: "Can computers do the same sort of things as humans can by way of their emotions; and can they be made to do so in a functionally similar way?" (Frijda and Swagerman, 1987, p. 236)

The starting point for ACRES is the adoption of a system that has a wide range of *concerns* and limited resources. In addition, this system moves in an environment that changes rapidly and is never completely predictable.

Based on these circumstances, Frijda and Swagerman define seven requirements for such a system:

1) The existence of concerns requires a mechanism that can identify objects with *concern relevance* - i.e. objects that can promote or prevent a concern.
2) Because opportunities and dangers are spread over space and time, the system must be able to act; otherwise, it cannot be considered independent. In addition, the action control system must be able to understand the signals of the *concern relevance* mechanism.
3) The system must have the ability to monitor its own activities in pursuing opportunities and averting dangers and to be able to recognize whether an action can lead to success or not.
4) The system must have a repertoire of appropriate alternative courses of action and be able to generate sequences of actions or plans.
5) The system requires a series of pre-programmed actions for emergencies so that it can respond quickly if necessary.
6) Since the environment of the system consists partly of other agents like it, actions with a social character must be present in the repertoire of actions.
7) Multiple concerns in an uncertain environment make it necessary to rearrange goals or temporarily put them on the back burner. The system must have a mechanism that allows for such changes in priority.

All these specifications, according to Frijda and Swagerman, are fulfilled by the human emotion system:

1) Objects are perceived as attractive or repulsive.
2) The attainment or non-attainment of such objects triggers signals of joy and pain.
3) Different emotions are responses to different situations with different relevance and include action selection processes.
4) Different emotions trigger different impulses and thus different plans of action.
5) Some of these emotion-action pairs are pre-programmed.
6) Social actions are a particularly prominent group of emotional actions.
7) Emotional action programs dominate non-emotional programs of action - and thus lead to the interruption of ongoing activities and the reorganization of the goals of the system.

In order to make it possible to implement a system that meets these specifications, Frijda and Swagerman chose interaction with the user of this program as an action environment that makes sense for a computer program. The concerns of the system in this context are:

- further work (*avoid being killed concern*);
- continuous work (*preserve reasonable waiting times concern*);
- Obtaining correct input (*correct input concern*);
- Receiving interesting input (*variety in input concern*);
- further work without changes (*safety concern*).

All knowledge of ACRES is organized in the form of concepts. These concepts consist of attribute-value pairs. Concerns are represented by a concept that contains the topic on the one hand, and on the other hand has a tariff (*tariff*) sub-concept that represents the desired situation.

ACRES has three main tasks: to receive and accept input (for example, the system rejects inputs with typos); to learn about emotions through information about emotions it receives from the user, and to collect, store and use knowledge about one's own emotions and those of others. Therefore, the system has three corresponding task components: *Input*, *Vicarious knowledge* and *Learning*.

Each task component has two functions: an *operation function* and a *concern realisation function*. The functions test whether concepts exist that are applicable to the information obtained; they use their knowledge to infer and generate related goals; they infer which actions are relevant to achieving these goals and trigger corresponding actions.

The essential information that ACRES works with results from the user's inputs, from information already collected by ACRES, and from information inferred by ACRES from the existing inventory. The information collected forms the "memory" of ACRES. This includes, for example, how often a particular user made typing errors; how long ACRES had to wait for new input, etc. Based on his experience with the users, ACRES creates a so-called *status index* for each user: positive experiences lead to a status increase, negative ones to a status reduction.

Concern relevance tests in ACRES are carried out in such a way that the information about a current situation is compared with the pre-programmed concerns of the system. In addition to the information that ACRES collects over time, there are some specific inputs that are directly emotionally relevant to ACRES, for example, the "*kill*" command.

Information about alternative courses of action is also represented in concepts in ACRES. Each action concept consists of the subconcepts *start state*, *end state* and *fail state*. The *subconcept start state* describes the starting conditions of an action, *end state* describes the state that the action can reach, and *fail state* describes the conditions under which this action cannot be executed.

In the selection of actions, the desired goal is first compared with the *end state* subconcepts of all action concepts; then the current state is compared with the *start state* subconcepts of the previously selected action concepts and one of them is selected. If there is no suitable action plan, a planning process is initiated, which first selects the action plan with the most obvious *start state*.

Events at ACRES lead to goals being set. The event of discovery of concern relevance leads to the goal of doing something in this regard. The following action selection process selects an alternative course of action according to the procedure described above. This process corresponds to what Frijda calls *context appraisal* in his emotion model .

Not every generated plan is executed in ACRES. This is ensured by *control precedence*. At ACRES, control priority means several things:

- time, processing capacity, and storage space are used to *concern realisation*goal. Task-oriented processing is postponed.
- Precedence remains in place if the user does not change the situation as a result of ACRES' requests.
- ACRES can refuse to accept new input unless its request has been realized.
- ACRES implements the *concern realisation*actions, some of which can influence the subsequent processing processes.

At ACRES, the primacy of control depends on two factors: the relative importance of the mobilized concerns and the gravity of the situation. The relative importance of the concerns is a fixed value; "*kill*" has the highest meaning of all. The seriousness of the situation is a variable that changes as ACRES interacts with users.

In order to be effective, the primacy of control must overcome a certain threshold.

The net result of all these processes is a series of "emotional" phenomena. For example, ACRES has a vocabulary of curses, insults or exclamations that can express such a state. The system may refuse to cooperate further with a user; can try to influence him or simply make the same demand to the user over and over again. The special thing about ACRES is not the fact that the program does not continue to work if the input is incorrect - every other software does the same:

> “It is the dynamic nature of the reactions, however, that is different: They sometimes occur when an input mistake is made or some other input feature is shown, and sometimes they do not. Moreover, some of the reactions themselves are dynamic, notably the changes in operator status.”
> (Frijda und Swagerman, 1987, p. 254)

In addition to perceiving events and their implications, ACRES is also able to perceive its own perception. The model constructs a representation of the current situation and the aspects relevant to its concerns. According to Frijda and Swagerman, ACRES constructs an emotional experience at the same time. They expressly emphasize: "It is not a play on words when we say that ACRES builds up an emotional experience." (Frijda and Swagerman, 1987, p. 254). They continue:

> “We do not wish to go into the deep problems of whether ACRES’ representations can be said to correspond to “experiences”, to “feels”, as they are called in philosophical discussion. Probably, ACRES cannot be said to “feel”, just as a colour-naming machine cannot be said to “see” red or blue, although we still have to be given satisfactory criteria for this denial of feeling or seeing. The main point, in the present context, is that ACRES shows many, and perhaps, in essence, the major signs that lead one to ascribe “emotions” to an animate agent.”
> (Frijda und Swagerman, 1987, p. 255)

The authors themselves admit that their model has a number of shortcomings. For example, ACRES is far from showing all the phenomena that occur in the attribution of emotions. However, they claim that these shortcomings are trivial from a theoretical point of view, since this is not a question of principle, but merely of implementation. They argue that the computer is not able to work in parallel and thus provide interruptions to an ongoing activity in real time. Furthermore, the computer does not move in a real environment and does not have its own energy balance. However, all these are purely technical questions and not problems of theory.

5.7. The Model of Moffat and Frijda

In another paper (Moffat, Frijda and Phaf, 1993), Frijda admits that ACRES does have some fundamental shortcomings. From this criticism, he developed, together with Moffat (Frijda and Moffat, 1993 and 1994), a new basic model that was rudimentarily implemented in the WILL system (Moffat and Frijda, 1995; Moffat, 1997).

5.7.1. Critical of acres

The starting point for the criticism of ACRES is a logged session with a user, which is typical for the system:

Hello, Dave. I am ACRES. What would you like to know?	
[1] show_concept X YOU KEEP ME WAITING TOO LONG!<< prints data about X >>	*Query the database about X.I had paused before typing. (details not important)*
[2] debug_on << prints lots of internal variables etc.>>	*Turn debugging on.*
[3] debug_off YOU KEEP ME WAITING TOO LONG!	*Turn it off again. I kept him waiting again.*
[4] debug_off YOU PUT TOO LITTLE VARIETY IN YOUR INSTRUCTIONS!	*I'm trying to `bore´* ACRES. *He's `bored´ with "debug_off".*
[5] debug_off YOU KEEP ME WAITING TOO LONG! YOUR INPUT IS CORRECT. FINE!	*I kept him waiting again. I made no typing errors yet.*
[6] debug_off YOU ENDANGER MY SAFETY! YOU PUT TOO LITTLE VARIETY IN YOUR INSTRUCTIONS!	*I ignore his requests to type faster, and to be more interesting, so he regards me as an unfriendly threat.*
[7] debug_off YOU PUT TOO LITTLE VARIETY IN YOUR INSTRUCTIONS! YOUR INPUT IS CORRECT. FINE!	
[8] debug_off YOU KEEP ME WAITING TOO LONG!	
[9] debug_off YOU ENDANGER MY SAFETY!	
[10] name_last_emotion YOU DO NOT KEEP ME WAITING. FINE! I experienced : Self-hatred ... 89 % Anger .... 88 % Disgust .... 79 %	*Ask the model to analyse itsemotional experience. He finds three that were quite strongly felt(100% is strongest).*

YOU ENDANGER MY SAFETY! YOU PUT TOO LITTLE VARIETY IN YOUR INSTRUCTIONS! YOUR INPUT IS CORRECT. FINE!	
[11] show_concept YOU KEEP ME WAITING TOO LONG! You did not supply an argument... which concept?	
[12] X << prints data about X again >>	
[13] kill YOU ENDANGER MY SAFETY! Please, do not let me die! Please type in "live"; it will kill me if you type "kill" again!	*First "kill" command to end session*
[14] debug_off Please, revoke the kill command by typing "live". YOU KEEP ME WAITING TOO LONG!	*Ignore request, but don't "kill" either.* *(Unusual aggression for someone making an appeal.)*
[15] live Thank you, may your life be very happy!	

Tab. 7: Sitzungsprotokoll von ACRES (Moffat et.al., 1993)

On the basis of this protocol, Moffat et al. formulate the following criticisms of ACRES:

1. ACRES shows enormous mood swings, and within a few seconds. One of the reasons for this, according to the authors, may be that ACRES does not have a physiology like humans, whose emotional states are given a certain stability and duration by neurotransmitters, for example. But it seems much more important to the authors that ACRES has no memory. Even a short-term memory, i.e. the ability to remember the state that preceded it, could influence the behavior of the system in a similar direction to physiology.
2. ACRES delivers contradictory emotional responses in a single output. If a user enters the same command over and over again, but quickly, ACRES will show a positive emotional response to the speed of the input, but a negative response to the lack of variability of the input. This is an atypical behavior for humans.
3. The emotional and unemotional responses issued jointly by ACRES do not concern the same topic, but different topics. Again, this is rarely observed in humans. ACRES can answer a user's question objectively and immediately afterwards output an emotional reaction to another point. The reason given by the authors is that ACRES cannot theoretically distinguish between emotional and more generally motivated behavior and considers this to be qualitatively equivalent. The reason for this lies in an arbitrarily set threshold value with which the system distinguishes between emotionally relevant and emotionally irrelevant concerns.
4. The reactions of ACRES are easily predictable. For example, input that is too slow is always answered with the phrase "You keep me waiting too long!". This corresponds more to a reflex than a real emotional reaction.

5.7.2. Emotional System Requirements

Based on this analysis, the authors then propose a number of other components that an emotional system should have and that also have implications for a theory of emotions.

Under the term awareness *of the present*, they describe the ability of a system to observe its own actions over a certain period of time. This *motivational visibility of the present* means that a system does not simply forget a motivated action that has failed, but that the emotion only disappears when the originally intended target state has been reached.

As a second necessary element, they cite the *motivational visibility of the planner*. In ACRES, as in almost all other AI systems, the planner is implemented as a separate module. The other modules do not get an insight into his half-finished plans and therefore cannot influence them. However, the various concerns of a system must be able to consult these plans, since they arise from very specific points of view which, although logical in themselves, may violate another concern.

The third element is referred to by the authors as *motivational visibility of the future* . This refers to the possibility of making not only one's own planned actions visible to the entire system, but also the actions of other agents and events from the environment. This is important for anticipations of the future and thus for emotions such as surprise.

Furthermore, the system requires a *world model*. In ACRES, only the planning module has such a world model. The overall system has no way of observing the effects of its actions and recognizing whether they have failed or been successful. Coupled with a memory, the world model gives the system the ability to try out and evaluate different actions. This gives the system a larger and, above all, more flexible repertoire of actions. At the same time, a *sense of time* is required, with which the system can estimate the time within which it must react and how long an action will take.

Finally, the authors consider it essential to clearly distinguish between motives and emotions, which ACRES does not do. They postulate that an emotion only occurs when a motive cannot be satisfied or can only be satisfied with a great strain on the resources of a system. A system will then first try to satisfy a request with the almost automatic action associated with it. If that doesn't work, or if the system can predict that it won't work, or if the system's confidence in it is low, or if the system assumes that it doesn't have enough control, then an emotion occurs. Their function is to mobilize the entire system to deal with the problem.

5.7.3. Implementation in WILL

Based on these explanations, Frijda and Moffat have developed a computer model called WILL, which is intended to remedy the shortcomings of ACRES. WILL is a parallel system with the following architecture:

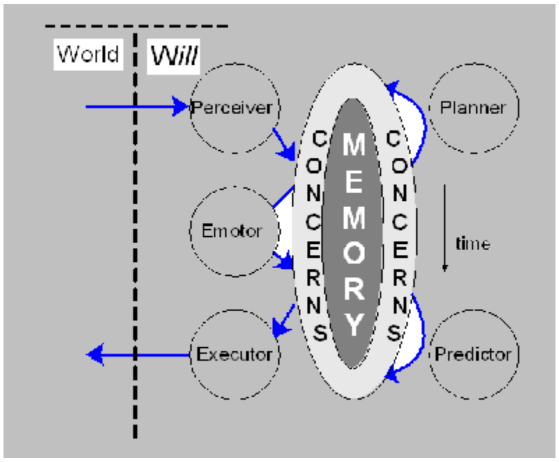


Fig. 5: Architecture of WILL (Frijda and Moffat, 1994)

The system consists of a perception module, the *perceiver*, an action execution module, the *executor*, a prediction module, the *predictor*, a planning module, the *planner*, and an emotion module, the *emotor*. It also has a central memory and a module for checking the relevance of concerns.

A basic principle of the system is that all modules do not communicate directly with each other, but only via the memory. This means that all elements of the system have access to all processes and sub-processes of other elements at all times. Each module reads its information from memory, processes it and writes it back to memory. The working methods of the modules are parallel, which means that they are all on an equal footing in principle.

Everything that is written to memory goes through a test for concern relevance when it passes through the *concerns layer*. This mechanism gives the system a regulatory authority, because different concerns have different meanings for the system. The Concern Relevance module therefore has a control function in that it evaluates the passing information differently.

This evaluation looks like this: the concern relevance module assigns a charge value to each element that is written to the memory. Elements with higher charge are more relevant to the concerns of the system than elements with low charge.

Each of the modules receives its information from memory. The element with the highest charge is always released to the modules to be processed by them. The authors call this element a focus item. To prevent the ranking of the elements in memory from remaining the same, the elements must be able to gain and lose charge. With WILL, this happens because the element in the memory with the highest charge loses charge if it is not processed by a module in a processing cycle. So if the *planner* has received a focus element but has not been able to develop a plan in connection with it, the element is written back to memory with a lower load. The authors call this method *autoboredom*.

The task of the *emotor* is to identify elements with high relevance to the concerns in the context of a further assessment process (Moffat calls it *secondary appraisal;* it corresponds to the *context* *appraisal* from Frijda's theory) to produce tendencies of action that belong to the emotion and to store them in the memory as intentions to act. In the next edit cycle, the *executor* will pick up on this action intention, unless it has been changed in the meantime or has lost its rank as a focus element.

Moffat has presented a first realization of WILL (Moffat, 1997). The system has the task of playing the game "Prisoner's Dilemma" with a user. In its basic form, Prisoner's dilemma consists in the fact that two game partners decide independently of each other whether they want to cooperate with their counterpart (*cooperate,c*) or not (*defect, d*). After they have made their choice, this two players will be announced. Depending on the result (*cc, cd, dc, dd*), players will be paid a certain amount of money. The scoreboard for WILL looks like this (numbers mean dollar amounts):

		User	
		c	d
Wants	c	3 3	5 0
	d	0 5	1 1

Table 8: Result table for Prisoner's Dilemma (according to Moffat, 1997)

Extensive studies (see e.g. Axelrod, 1990) have shown that in most circumstances a strategy of mutual cooperation is most successful for both sides. However, there may be situations where it is better for a player not to cooperate.

In Moffat's model, there are two types of events: *move events* and *payoff events*. The user's move is formally rendered with *move(user,c)*. A prediction of which move the user will make in the next round is expressed with *move(user, {c,d})*. This allows the world of the game to be expressed in a structured form. Thus, the assumption of WILL that he does not cooperate, but the user either cooperates or does not cooperate, is expressed with the associated rewards as follows:

move(will,d) & move(user, {c,d}) ==> payoff (will, {1,5}) & payoff (user, {0,1}

WILL's aim in this game is to win as much money as possible. This is formally expressed as *$_concern = [0 -> 2 ->5]* and means that the most undesirable result is 0 dollars, the most desirable 5 dollars and the so-called *set-point* 2 dollars. The *set-point* defines what the average result is. The valence of the possible outcomes for WILL is defined as follows:

win $0 --> valence = -2 x IMP
win $2 --> valence = 0 x IMP
win $3 --> valence = +1 x IMP
win $5 --> valence = +3 x IMP

IMP is a factor in the importance of the concern.

Another concern of WILL is moral behaviour. The system knows that cooperation is more moral than non-cooperation:

morality_concern = [0 -> 0.8 -> 1].

The move *c* has the moral value 1, the move *d* the value 0. The *set-point* is 0.8.

WILL has two cognitive modules, the *Predictor* and the *Planner*. A *world* model is implemented in *memory*, which expresses, for example, the assumption that the user constantly does not cooperate as follows:

move(user,UM) --> move(user,d).

With the elements mentioned, according to Moffat, essential parts of an emotion are already modeled, namely affect, relevance assessment and priority of control. The emotor is responsible for contextual assessment and the tendency to act . The assessments programmed into WILL are derived from Frijda's theory. Some examples:

Valence - Can be + or -. Says how un/ pleasant the emotion is.

Un/Expectedness - Was the perceived event expected?

Control - Does the agent have control over the situation?

Agency - Who is responsible for the event?

Morality - Was the event (the action) moral?

Probability - The probability that the event will actually occur.

Urgency - How urgent is the situation?

Action tendencies are also firmly programmed into WILL. Some examples:

hurt(O) / help(O) - Wants to harm or help another agent O.

try_harder(G) / give_up(G) – Push harder or give up on Goal G.

Approach(O) / Avoid(O) - Will and Nahekammen are their Farnbleiben.

fight(O) / flee(O) - Will fight O or flee from him.

exuberance / apathy & inhibition - General level of activation.

Based on the assessments and tendencies to act, the *emotor produces* emotions that Moffat describes as true. He gives three examples:

Happiness	Appraisals:	valence = positive
		agency = world

	Action tendency:	happy_exuberance

Anger	Appraisals:	valence = negative
		morality = negative
		agency = User

	Action tendency:	hurt(User) --> play D, attend(User)

Pride	Appraisals:	valence = positive
		morality = positive
		agency = Self

	Action tendency:	exuberance --> verbal, attend(Self)

Using the minutes of the meeting, Moffat then explains the internal workings of WILL:

1. Planner: Notice that I play **c** or **d** in round 1.
a. Decide that I play **c** in round 1.

WILL has noticed that he is about to play a first round of Prisoner's Dilemma. The *planner* shows the two alternatives; the decision falls on *c*, because this is the morally more correct alternative.

2. Predictor: Notice that I play **c** in round 1.
a. Predict that I win $0 or $**3** and User wins $3 or $5 in round 1.

The *predictor* picks up the information written back to the memory and predicts the possible results of the first round.

3. Predictor: Notice that I win $0 or $3 in round 1.
a. Predict that I play **c** or **d** and User plays **c** or **d** in round 2.
4. Predictor: Notice that I play **c** or **d** in round 2.
a. Predict that I and User win $0 or $1 or $3 or $5 in round 1.

The *predictor* reads the information again and makes further predictions.

5. Planner: Notice that I play **c** or **d** in round 2.
a. Decide that I play **c** in round 2.

The *planner* reads the information and plans for the second round.

6. Executor: Tell the Umpire that I play **c** in round 1.

The *executor* performs the action proposed by the Planner for the first round and reports it to the umpire, a software module independent of the system. The perceptible change of theme illustrates how the system's attention shifts through the charging or discharging of elements in the memory: For several work cycles, the train for round 1 was so little charged that the other modules did not deal with it.

7. UMPIRE: Round 1. What do you play ? . . . **c**.
8. UMPIRE: Round 1. You play **c** and Will plays **c**.
9. Perceiver: Hear from Umpire that User just played **c** and I just played **c**.
10. Emotor: Notice that I just played **c** in round 1.
a. **Appraisals**
b. intensity = 0.4 **Action tendencies**
c. valence = +0.4 exuberance = 0.4
d. agency = myself
and. morality = 0.4 **emotion** is *pride*
f. [0.4] express pride

The referee announces the moves of the first round. The *perceiver* writes them to memory. The *emotor* perceives it and develops a positive emotion due to its moral move, the value of which, however, lies below the arbitrarily set threshold for verbalization.

11. UMPIRE: Round 1. You win $3 and Will wins $3.
12. Perceiver: Hear from Umpire that User just won $**3** and I just won $3.
13. Emotor: Notice that I did win $3 in round 1.

	a.	**Appraisals**	
	b.	intensity = 4.0	**Action tendencies**
	c.	valence = +4.0	jump_for_joy = 4.0
	d.	agency = world	
	and.	morality = 0.0	**emotion** is *joy*
	f.	[4.0] express joy	
	say:	*"La la la!"*	

The referee announces the winnings of the first round. The *perceiver* writes them to the memory, where they are perceived by the *emotor* , which then produces an emotion with a sufficiently high value for it to lead to verbalization.

14.	Emotor:	Notice that I did win $3 in round 1.	
	a.	**Appraisals**	
	b.	intensity = 2.7	**Action tendencies**
	c.	valence = +2.7	jump_for_joy = 2.7
	d.	agency = world	
	and.	morality = 0.0	**emotion** is *joy*
	f.	[2.7] express joy	

Since no other module works with the information of the *emotor* , a discharge process takes place (*autoboredom*). The value of emotion decreases.

15.	Emotor:	Notice that I shall win $0 or $1 or $3 or $5 in round 2.	
	a.	**Appraisals**	
	b.	intensity = 3.0	**Action tendencies**
	c.	valence = +3.0	jump_for_joy = 6.0
	d.	agency = world	
	and.	morality = 0.0	**emotion** is *joy*
	f.	[6.0] express joy	
	say:	*"Yabba-dabba-doo!"*	

The *emotor* reads out the winning expectations for round 2 and develops a corresponding anticipation with a high value.

. . .

16. UMPIRE: Round 2. You play **d** and Will plays **c**.

. . .

17.	Emotor:	Notice that User just played **d** in round 2.	
	a.	**Appraisals**	
	b.	intensity = 1.8	**Action tendencies**
	c.	valence = -1.8	sentiment = -2.7
	d.	agency = user	so urge = 4.5 (|int-sent|)
	and.	morality = -1.8	hurt(user) = 4.5
	f.	[4.5] express anger	**emotion** is *angry revenge*
	say:	*"I will get you back for that!"*	& soon play **d** to hurt user

(Several intermediate steps are omitted.) The referee announces the moves of round 2. The user's turn *d* means that WILL goes away empty-handed. This angers WILL because it not only violates his moral standards, but also interferes with his purpose of making money. The value of the emotion produced by the emotor is correspondingly high . This triggers the tendency to play d in the next round to get back at the user.

In a subsequent discussion, Moffat asks whether WILL has a personality as defined by the five big traits (*big traits*). He states that WILL is neither open nor agreeable: he has too few interests for that and has no social consciousness. However, he was neurotic, because WILL was a *worrier*; moreover, it could be said that he had, at least in part, a conscience (*conscientious*) - because he was endowed with a concern for fairness and honesty. The characteristic of extroversion could also be attributed to him in part. Moffat comes to the conclusion that machines can certainly possess human-like personalities:

> "In this case, the answer is a much more qualified "yes"... The programmable personality parameters in Will include the charge manipulation parameters in the attentional mechanism, the appraisal categories, action tendencies, and concerns, all of which can be set at different relative strengths. In this programmability, human-specificity can be built in as required, but with different settings other personalities would also be possible, the like of which have never yet existed. What they may be is beyond science-fiction to imagine, but it is unlikely that they will all be dull, unemotional computers like HAL in the film 2001."
> (Moffat, 1997)

5.8. Other Models

There are a number of other models that deal with the simulation of emotions in computers from different aspects. Some of them will be briefly touched upon in this section; a more detailed discussion would go beyond the scope of this work.

5.8.1. Colby's Model

One of the first computer models that explicitly dealt with emotions was PARRY by Kenneth Colby (Colby, 1981). PARRY simulates a person with a paranoid personality who feels persecuted by the Mafia. The user interacts with the program in the form of a dialog in which the system responds to text input via the keyboard with verbal output.

The program is tasked with actively searching the user's input for an interpretation that can be interpreted as malice. As soon as this is discovered, one of three emotional reactions is triggered: fear, anger or mistrust, depending on the type of imputed malice. An assumed physical threat triggers fear; an assumed psychological threat triggers anger; both types of assumed threat also trigger mistrust. PARRY reacts to the attacks he has constructed either with a counterattack or with a retreat.

In order to construct a model of a paranoid patient, Colby and his colleagues invested a lot of work in the project by the standards of the time. PARRY has a vocabulary of 4500 words and 700 colloquial phrases, as well as the grammatical competence to use them. PARRY compares users' text inputs with its stored list of words and phrases and reacts in emotional mode as soon as it detects a match.

A number of variables relate to the three emotional states of fear, anger and mistrust and are constantly updated in the course of an interaction. This allows PARRY to "get into" certain emotional states more and more; even Colby, his creator and a psychiatrist by training, was surprised by some of PARRY's behaviors.

Colby subjected PARRY to several tests. In one, he had several psychiatrists conduct interviews with paranoid patients and with PARRY by telex, without informing them that a "patient" is a machine. After their interviews, Colby explained this to the participants and asked them to identify the machine. The result: Apart from one or two random hits, no psychiatrist could say whether he had talked to a person or to PARRY.

In another experiment, an improved system was again presented to a number of psychiatrists. This time, the test participants were informed in advance that one of their participants in the conversation would be a computer, and they were asked to identify him. Once again, the result was not much different from the first time.

PARRY has the possibility to express *beliefs*, *fears* and *anxieties*, but these are predefined and fixed from the outset. Only the intensity can change in the course of an interaction and thus modify the conversational behavior of PARRY.

5.8.2. The Reeves Model

THUNDER stands for *THematic UNDerstanding from Ethical Reasoning* and was developed by John Reeves (Reeves, 1991). THUNDER is a system that can understand stories and focuses on the evaluation of these stories and ethical considerations.

In order to represent different points of view in a conflict situation, THUNDER uses so-called *Belief Conflict Patterns*. This enables the system to work out moral patterns from presented stories. These patterns, in turn, are used by so-called evaluators to make moral judgments about the characters in a story. Without such moral patterns, according to Reeves, many texts (and situations) would not be understood.

As an example, Reeves cites the story of hunters who tie dynamite on a rabbit "for fun". The rabbit hides with the dynamite under the hunters' wagon, which is destroyed in the following explosion. To understand the irony of such a story, Reeves argues, the system must first know that the actions of the hunters are morally reprehensible, and that the subsequent, accidental destruction of their car is a morally satisfying compensation.

The focus of THUNDER is on the analysis of motives of other individuals who are either in a certain situation or observe it.

5.8.3. The Model of Rollenhagen and Dalkvist

Rollenhagen and Dalkvist have developed SIES, the *System for Interpretation of Emotional Situations* (Rollenhagen and Dalkvist, 1989). The task of SIES is to draw conclusions about situations that have triggered an emotion.

SIES combines a cognitive with a situational approach. The basic assumption is that the triggering conditions of an emotion are to be found in situations of the real world. The core of SIES is a *reasoning system*, which carries out a structural content analysis of submitted texts. These texts consist of reports in which emotion-triggering situations are reported retrospectively.

The system is equipped with a set of rules that, while capable of differentiating and classifying emotions, ultimately do nothing more than structure the information contained in a story.

5.8.4. The Model of O'Rorke

The AMAL system presented by O'Rorke and Ortony (O'Rorke and Ortony, unpublished manuscript), later referred to by Ortony as "AbMal", is based on the theoretical approach of Ortony, Clore and Collins (1988). The goal of AMAL is to identify emotion-triggering situations that are described in student diaries. To solve this task, AMAL uses a so-called *situation calculus*. With the help of abductive logic, AMAL can filter out plausible explanations for their occurrence from emotional episodes.

5.8.5. The Model of Araujo

Aluizio Araujo from the University of Sao Paulo in Brazil has developed a model that attempts to combine findings from psychology and neurophysiology (Araujo, 1994).

Araujo's interest lies in the simulation of mood-dependent memory, learning and the influence of anxiety and task difficulty on memory performance. His model consists of two interacting neural networks, the "emotional network" and the "cognitive network". The intention is to simulate the roles of the limbic and cortical structures in the human brain. For Araujo, it is essential to model not only cognitive processes, but also physiological emotional responses at a low level, which influence cognitive processing at a higher level.

The emotional network evaluates affective meanings of incoming stimuli and produces the emotional state of the system. Its processing mechanisms are relatively simple and therefore fast. The cognitive network performs cognitive tasks, such as freely remembering words or associating between word pairs. The processing processes are more detailed than in the emotional network, but also require more time.

In Araujo's model, an "emotional processor" calculates valence and arousal for each stimulus and thereby changes parameters of the cognitive network. In particular, the output of the emotional network can influence the learning rate and accuracy of the cognitive network.

5.9. Summary and evaluation

The models presented in this chapter differ significantly in their theoretical prerequisites and in their details. Nevertheless, there is one thing in common: The goal of all models is either to understand emotions or to exhibit (pseudo-)emotional behavior. What an emotion is is precisely defined from the outset. The differences between the models lie mainly in the variety of defined emotions as well as in the richness of detail of the model.

The approaches of Elliott and Reilly are based on the emotion theory of Ortony, Clore and Collins. Their goal is to increase the performance of a computer system in certain tasks by taking emotions into account, for example in speech understanding or in the development of computer-aided learning systems or other interactive systems. In these models, the introduction of emotional elements is carried out on the basis of predetermined tasks that need to be fulfilled better. Both Elliott and Reilly have fulfilled a large part of the claim they set themselves with their systems. However, it becomes clear that the operationally formulated theory of Ortony, Clore and Collins cannot simply be translated into a computer model, but must be extended by additional components whose significance within the framework of the theory is doubtful. In particular, Reilly's critique of the "over-cognitivization" of the theory has led him to introduce a "shortcut" that is not simply an extension of the OCC model, but stands outside of it.

Dyer's BORIS and OpED models, as well as AMAL, THUNDER and SIES, only serve to identify emotions from texts, but perform this task less efficiently than Elliott's *Affective Reasoner*.

In contrast, the approaches of Frijda and his colleagues pursue the goal of testing Frijda's theory of emotion using a computer model. The deficits that occurred at ACRES have led Frijda to a partial revision of his theory, which will now be re-examined in WILL. The models have no other task than the implementation and testing of the theory. The same applies to GENESE by Scherer, although the level of detail of this model is much lower than, for example, WILL.

DAYDREAMER by Mueller and Dyer and WILL by Frijda and Moffat are already on the verge of models in which emotions are seen as control functions of an intelligent system. Pfeifer's FEELER was also born out of the claim to simulate control processes; however, the model is not able to do so due to its very specific definitions of emotions.

Finally, Colby's model has more historical interest, as he was less interested in modeling emotions and more interested in simulating a specific phenomenon that included an emotional component.

6. The Pioneer’s Visions

Herbert A. Simon began his academic career in the 1940s as a professor of political science before taking over the chair of computer science and psychology at Carnegie-Mellon University in 1949, which he still holds today. His versatility is also evident in the fact that he received the Nobel Prize in Economic Sciences in 1978 - although formally he has nothing to do with this academic discipline. But Herbert Simon has always been a person who has crossed boundaries.

He is not necessarily someone who hides his light under a bushel. Asked about the "cognitive revolution", he replied succinctly in an interview: "You might say that we started it." (Baumgartner and Payr, 1995, p. 233)

By "we" he meant himself and his partners, Alan Newell and J.C. Shaw. Together with them, Simon had developed a computer program called "Logic Theorist" (LT) between 1955 and 1957, which was supposed to prove theorems by heuristic search. LT evolved into GPS, the *general problem solver* developed by Newell, Shaw and Simon between 1957 and 1959.

The GPS was the first computer program that had been explicitly developed to simulate human problem-solving processes. In doing so, Simon and his colleagues broke new ground at a time when behaviorism dominated. At the same time, they laid the foundation for a series of further attempts to understand the functioning of the human mind with the help of computers.

In 1967, Herbert Simon published an essay in the Psychological Review entitled "Motivational and Emotional Controls of Cognition" (Simon, 1967). In it, he made emotions part of a systematic modeling approach of cognitive processes for the first time.

The work was written in response to an article by Ulric Neisser. In it, Neisser expressed his criticism of computer programs that existed or were planned at the time as follows:

> ”Three fundamental and interrelated characteristics of human thoughts... are conspicuously absent from existing or contemplated computer programs:
> 1) human thinking always takes place in, and contributes to, a cumulative process of growth and development;
> 2) human thinking begins in an intimate association with emotions and feelings which is never entirely lost;
> 3) almost all human activity, including thinking, serves not one but a multiplicity of motives at the same time.”
> (Neisser, 1963, p.195; quoted from Simon, 1967)

Simon accepted Neisser's objections and understood his work as an attempt to lay a first theoretical foundation for the construction of an information-processing system that has emotions and multiple goals.

Neisser and other critics of computer modeling of mental processes had pointed out, among other things, that they have little to do with human behavior. For example, such programs would only pursue a simple goal and not, like humans, be driven by numerous motives.

For Simon, this argument was not valid. He admitted that the implemented models were "excessively simplified" (Simon, 1967, p. 34); however, this was due to technical requirements. However, the underlying models of hierarchically structured, serial information processing are not so one-dimensional:

> ”Activity towards specific goals is terminated by aspiration, satisficing, impatience, and discouragement mechanisms; distinct tasks may be queued or handled within individual time allocations; choices among alternatives may respond to multiple criteria.”
> (Simon, 1967, p. 34)

At the same time, however, Simon was also aware that such a model has shortcomings:

> ”The mechanisms we have considered are inadequate to deal with the fact that, if the organism is to survive, certain goals must be achieved by certain specified times.”
> (Simon, 1967, p. 34)

What the previous models lack is clear: a mechanism that can "hijack" attention at any given time in order to make it usable for goals that are essential for survival. "If real-time needs are to be met, then provision must be made for an *interrupt system*." (Simon, 1967, p. 34)

Simon then develops a theory of such an interrupt system. First, he defines three classes of real-time needs of an individual. *Needs arising from uncertain environmental events* are, for example, sudden noises or visual stimuli that could signal danger. *Physiological needs* are internal stimuli that indicate physical needs, for example hunger, thirst, exhaustion, etc. *Finally, cognitive associations* are strong stimuli that are triggered by memory associations, for example an unspecified fear.

These real-time needs, according to Simon, are accompanied by a number of physiological phenomena as well as subjective feelings, which commonly accompany the states that are referred to as "emotion".

Such *an emotional stimulus* fulfills an essential survival function as an interruptor by interrupting ongoing processing processes and drawing attention to a problem that is more urgent for the survival of the individual. Under certain circumstances, however, the *interruptor* can also become a *disruptor*, which then no longer has any adaptive value.

An essential quality of the *interrupt system* is that it can be changed through learning processes.

> ”In two ways, then, we may expect learning to reduce the emotionality of response as a situation becomes more familiar: (a) The need for interruption is reduced by incorporation of more elaborate side conditions in the programs associated with ongoing goals; (b) the response to interruption becomes more successfully adaptive, thus forestalling new interruptions.”
> (Simon, 1967, p. 37)

As a conclusion of his considerations, it is clear to Simon that realistic and promising theories of human cognition must include emotions in the form of an *interrupt system* .

Simon summarizes his theory as follows:

> “The theory explains how a basically serial information processor endowed with multiple needs behaves adaptively and survives in an environment that presents unpredictable threats and opportunities. The explanation is built on two central mechanisms: 1. A goal-terminating mechanism [goal executor]... 2. An interruption mechanism, that is, emotion, allows the processor to respond to urgent needs in real time.”
> (Simon, 1967, p. 39)

The theory implies that organisms have two parallel processing systems: a "goal executor" that generates actions, and an "observation system" that continuously checks the internal and external environment of an organism to see if an event requires a rapid response. The first, resource-limited system can be interrupted by the second.

With his work, Simon has defined a number of essential cornerstones that are important for the further development of autonomous systems. Such systems are driven by different motivations that can arise due to changing external or internal states. Due to the fact that such systems have limited resources but operate in a complex and largely unpredictable environment, they need a system of control structures that enables them to interrupt ongoing processes and initiate new ones when this is important for the survival of the system.

Simon deliberately limits his considerations not only to humans or animals, but considers them as design requirements for any autonomous system. So it is certainly no coincidence that its central mechanism is the interrupt, a term that is also used in a similar form in computer science.

Sloman (1992) explicitly interprets Simon's remarks as instructions for the construction of autonomous systems:

> “He outlines some of the control issues, and suggests suitable mechanisms, inspired in large part by developments in computer science and AI, including software techniques for generating new sub-goals at run time, techniques for queueing and scheduling processes, techniques for forming plans in order to achieve goals, techniques for assigning priorities and resolving internal conflicts, and techniques for generating and handling interrupts.”
> (Sloman, 1991, p. 12)

7. Encounters on Taros

One of the most important impulses for the modeling of emotions in the computer comes from the Japanese psychologist Masanao Toda. It is the design of an autonomous robot system, the so-called Fungus Eater.

Masanao Toda was born in 1924 in Okagi, Japan. After graduating from school, he studied physics at the Imperial University of Tokyo. After the war, he worked as a mathematics and physics teacher at a secondary school and began studying psychology at the University of Tokyo in 1949. A few years after graduating, he took up a chair in psychology at the University of Hokkaido.

Toda brought to experimentally oriented psychology a sharp mind accustomed to deductive thinking from theoretical physics. Although he worked extensively experimentally, his basic philosophy was:

> "... what finally counts are theories and ideas, no matter where they were originally hatched, either in an armchair or in an experiment. If an idea is good, it will eventually find a way to be experimentally tested, while a blind experiment produces only a trickle of possible facts out of the whole ocean of possibly obversable facts."
> (quoted by Hans F.M. Crombag in Toda, 1982, p. XIV)

Even in the 50s, the heyday of behaviorism, Toda could not make friends with this school of thought. For him, behavior was always the result of a personal choice between several possible alternative courses of action. He saw the psyche as a mediator between the demands of the environment and actions. In this respect, Masanao Toda was already a kind of cognitivist back then. Behaviorist was only his basic assumption that the human psyche and human behavior are responses to the demands of the environment.

Between 1961 and 1980, Masanao Toda developed his theory of the Fungus Eater; the corresponding essays were collected in his book "Man, Robot, and Society" in 1982.

7.1. What is a Fungus Eater?

The model of the Fungus Eater resulted from Toda's dissatisfaction with experimental psychology.

> "Psychology... will tell you a lot about human beings in experimental laboratories. Experimental laboratories are, however, not our natural habitat. The major difference between these two types of environment can be stated this way: In experimental laboratories, information is usually coded in a single - or at most, a couple of - sensory dimensions in a fairly abstract way, and the kind of task given to human subjects usually requires persistent, single-track thinking......
>
> But... human beings handle multiple-channel information input efficiently and engage in multiple-track thinking in their natural habitat. ...
>
> ... Experimental psychology tells us facts, but to obtain these facts we have been sacrificing important information coming from the multiplicity of our input channels and the multiplicity of our thinking and other activities....."
> (Toda, 1982, p. 94)

Based on this criticism, Toda first developed the Fungus Eater as the main character of an experimental situation in which participants play a kind of science fiction game. Perception, learning, thinking, behavior and the effective organization of these activities were simultaneously challenged and should result in a better experimental situation.

The Fungus Eater was described to the test subjects as follows:

> "You are a remote control operator of the robot miner nicknamed "Fungus-Eater", sent to a planet called Taros to collect uranium ore, which uses wild fungi growing on the surface of the planet as the main energy source for its biochemical engine. The uranium ore and fungi are distributed over the land of Taros, which is covered mainly with black and white pebbles, and little is known about the mode of their distribution. As the operator you can control every activity of the Fungus-Eater, including the sensitivity of the fungus- and uranium-detection devices. All the sensory information the robot obtains will be transmitted here and displayed on this console so that you will feel as if you are the Fungus-Eater itself.
> Note that your mission is to collect as much uranium ore as possible, and your reward will be determined in proportion to the amount of uranium you collect. Note also that the amount of fungi you collect and consume during your mission is irrelevant to the reward. Remember, however, that every activity of the Fungus-Eater, including the brain-computer operations, consumes some specified amount of fungus-storage. Never forget that the Fungus-Eater cannot move to collect further uranium ore or fungi once it runs out of its fungus-storage, and your mission would be over then and there. Good luck!"
> (Toda, 1982, p. 95)

What at first glance looks like a simple role-playing game turns out to be a situation that results in highly complex behavior on closer inspection. One is reminded of the "vehicles" of Braitenberg (1993), whose behavior, which the observer considers to be "complex", is the result of a few simple rules.

On the one hand, the Fungus Eater has a rudimentary system of attention control. Once he has consumed enough nutrients, he can concentrate fully on collecting ore and vice versa.

On the other hand, it has a system of different goals. His mission is to collect as much ore as possible; for this purpose, however, it must replenish its nutrient supply again and again. This construction can lead to conflicting goals, in which the fungus eater has to weigh up whether to collect ore or mushrooms based on different criteria.

Such a decision situation is still relatively trivial if the Fungus Eater only has to decide between two alternatives (ore or mushrooms) at a given time, i.e. if he is at a point on Taros from which he can locate an ore deposit on the right and a mushroom deposit on the left, for example. However, as soon as other factors such as obstacles, changing lighting (day/night) etc. are added, the Fungus Eater has to plan for the longer term. This complicates the model, because "thinking" also costs energy, which is thus lost for collecting ore.

Another factor, which in turn has serious consequences, is the assumption that there is not one, but several fungus eaters on Taros. This presents the system with completely new challenges, which make the decision problems of the solitary fungus eater seem almost trivial.

These few explanations make it clear that even a few simple basic assumptions can produce complex planning and decision-making processes that are not explicitly formulated in the basic model.

7.2. Emotional Fungus Eater

In a further thought experiment, Toda speculated about the consequences for his model if the Fungus Eater had emotions. For him, emotions are a mandatory prerequisite for the survival of a humanoid robot:

> "My intention is to demonstrate that a group of experimental humanoid robots, sent to some biologically wild environment, would have to be programmed to be more emotional than intellectual in order to survive there."
> (Toda, 1982, p. 130 f.)

Toda calls emotions urges in his model. Pfeifer (1988) sees a connection between Todas *urges* and Frijdas *concerns* insofar as:

> "... that urges are the programs which are activated once a situation has been identified as being relevant to some concern."
> (Pfeifer, 1988, S. 305)

Toda defines an *urge* as a built-in motivational subroutine that links cognition to action.

> "A separate set of cognitive contents is responsible for the activation of each urge, while each member of the set is characterized by a value corresponding to its estimated *relevance* to the issue of survival. Whenever one of the members of this set is brought into cognition, the urge subroutine is activated or "called", with the relevance value of the cognition transferred as the urge intensity, and the subroutine will be immediately executed if no competing urge with a higher intensity exists."
> (Toda, 1982, p. 136)

The importance of such a cognitive element for the current behavior of the fungus eater is determined by two variables: first, past experiences, i.e. learning; second, the context in which the fungus eater currently finds itself. This context dependency is controlled by a mechanism called toda *mood control* .

The *mood control* with its associated *mood operators* determines the importance attached to cognitions, thus acting as a kind of threshold setting. For example, the news of the death of another Fungus Eater by an enemy will cause the *Startle Urge* of the other Fungus Eater to be activated by even the slightest changes in perception.

Toda classifies his *urges* into four broad groups: "biological urges", "emergency urges", "social urges" and "cognitive urges".

7.2.1. The "biological urges"

Biological urges are primarily concerned with maintaining good physical condition and, according to Toda, are relatively independent of each other. Their main characteristics are similar to those of *emergency urges*, but usually with a much lower level of arousal.

Biological urges include elementary needs, for example *hunger urge*, although it is already questionable here whether the equation of *urges* with emotions is justified.

7.2.2 The "emergency urges"

One of the *emergency urges* is Toda

- *Startle Urge*
- *Fear Urge*
- *Anxiety Urge*

These three are not independent of each other, but are closely related.

The *Startle Urge* is activated whenever an unexpected stimulus is detected in the environment of the Fungus Eater and leads to the initiation of three parallel processes:

1) stopping all ongoing actions;
2) physical arousal;
3) concentrated cognitive effort to identify the source of the disorder.

In other words, the *Startle Urge* leads to cognitive information processing, attention control, and physical arousal. If the third process actually identifies a threat, the *Fear Urge is* initiated.

This is where Toda brings two more parameters into play: *intensity* and *importance*.

> “The cognition that has activated an urge will also determine the intensity of the urge, depending mainly on the appraisal of the importance of the urge activities in relation to the survival or welfare of the individual.... Once so determined, the intensity of an urge will function as the urge-regulating parameter.”
> (Toda, 1982, p. 134)

This design allows the Fungus Eater to have competing *urges* and give priority to the most important one, because the *urge* with the highest intensity controls the behavior.

If the Fungus Eater cannot identify a direct source of danger after the *Startle Urge*, it initiates the *Anxiety Urge*, which is characterized by a constant shift of attention from one potential source of danger to the next.

Each *urge* has a predefined set of action instructions. The result of the three processes mentioned above is the selection of a specific *action plan* from this repertoire.

7.2.3. The "social urges"

Social urges are important to Fungus Eaters because they help them have a cooperative social life. It is important to know that Toda's Fungus Eater Society is a hierarchically structured system. Toda groups his *social urges* into three categories:

a) *Helping urges*
Rescue Urge, Gratitude Urge, Love Urge

b) *Social System urges*

Protection Urge, Demonstration Urge, Joy Urge, Frustration Urge, Anger Urge, Grief Urge, Hiding Urge, Guilt Urge

c) *Status-related urges*
Confirmation Urge

At this point, the definitions of the individual *social urges* will not be discussed in more detail. It should only be noted that here, too, a very complex social interaction results from relatively simple basic elements that are given to the Fungus Eater.

7.2.4. The "cognitive urges"

Unfortunately, Toda's remarks on cognitive *urges* are only very incomplete, since social *urges* are of much greater importance for his model. He leaves it at explicitly naming only one *cognitive urge* , the *Curiosity Urge*.

The definition of what constitutes a *cognitive urge* can be found elsewhere (Toda, 1982, p. 151). There, but in a completely different context, Toda defines a *cognitive urge* as a learned *a posteriori urge*, which he also refers to as a motivational process.

7.3. Evaluation of Toda's model

The significance of Toda's model lies primarily in the fact that his Fungus Eater is an autonomous being who has to survive in an uncertain and unpredictable environment and is unable to do so without emotions.

Although the Fungus Eater has never been implemented by Toda in an actual computer model, it has all the prerequisites for it. Toda himself made initial proposals for operationalization in a paper entitled "The Design of a Fungus-Eater" (Toda, 1982).

Furthermore, the model impressively demonstrates the complexity that can develop in a system that only has a number of simple basic functions. It is this "emergent" behavior that makes Toda's model relevant again today.

Pfeifer also sees the importance of the fungus eater in its epistemological dimension:

> "The Fungus-Eater clearly becomes emotional when the necessary mechanisms are introduced for functioning in a "wild" environment, for which the human emotional system was obviously originally designed."
> (Pfeifer, 1988, S. 305)

It is noticeable that Toda is quite generous with the concept of *urges* . This is also confirmed when you look at his tentative proposals for *Rule Observance Urges* or *Ambition Urge* . These *urges*, which he equates with emotions, are largely defined a priori by him. One reason for this arbitrary approach may be that he did not translate his theory into an actual computer model and thus observe which emotions would develop due to the interaction of the few basic parameters.

In this respect, Toda's model in all its details is certainly not a useful model for modelling emotions; but his basic principles of an emotional autonomous agent are.

Especially under the aspect that artificial intelligence has almost completely neglected this aspect of modeling in recent decades, the heuristic value of Toda's model cannot be overestimated.

8. Further Development and Implementation of Toda's Model

In recent years, interest in Toda's theoretical approach has been reawakened. This is also reflected in the frequency with which he is quoted approvingly, for example by authors such as Frijda, Pfeifer or Dörner.

The increased reception of Toda coincides with an increased interest in the construction of real autonomous agents. In this context, there have been several approaches to modify Toda's model based on current results from emotion psychology and to partially implement it in a computer or robot simulation. The work of Aubé, Wehrle, Pfeifer and Dörner et al. is briefly presented below.

8.1. Aubé's modification of Todas *urges*

Michel Aubé has pointed out the problems of the system of Todas *urges* (Aubé, 1998). On the one hand, he criticizes Toda's classification of *urges*: For example, *grief* is found there among the *rule-observance urges*. On the other hand, he notes that Toda counts a number of *urges* among the emotions that would rather be described as need (e.g. hunger). Finally, he notes that some of the *urges* represent what Frijda calls *action tendencies* rather than the emotions themselves, such as *rescue* or *demonstration*.

Aubé therefore proposes to refrain from defining *the urges* as emotions for the time being, but rather to understand them as motives. Aubé differentiates these motives into two classes: Needs such as hunger or thirst represent a motivational control structure that enables access to and management of first-order resources. Emotions such as anger or pride are motivational control structures that create, promote or protect second-order resources. For Aubé, such second-order control structures are social commitments.

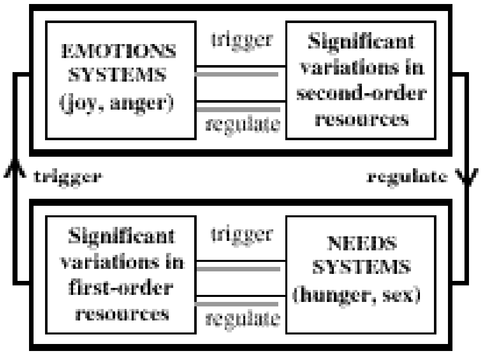

EMOTIONS SYSTEMS (joy, anger)
trigger
regulate
Significant variations in second-order resources
trigger
regulate
Significant variations in first-order resources
trigger
regulate
NEEDS SYSTEMS (hunger, sex)

Figure 6: Two layers of control for the management of resources (Aubé, 1998, p. 3)

For Aubé, commitments are the central factor in emotions:

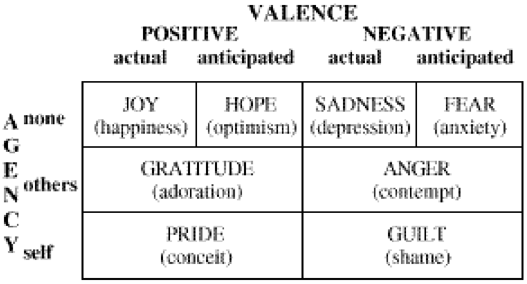
> “Since emotions in our model are specifically triggered as handlers or operators whenever some commitment is seriously at stake, we see commitments as the critical hidden variable behind any emotional reaction, just as the concept of force

> in physics is understood as the general cause to be invoked whenever a change in motion is observed."
> (Aubé, 1998, p. 4)

Within autonomous agents, *commitments* represent dynamic units, active subsystems that look for significant events that are important for their fulfillment or violation. They register variables such as who *is beholden* to whom, until *when* and why (*about what*).

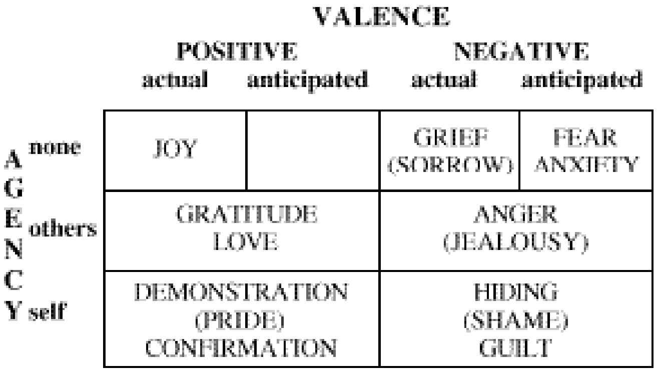
> "«To whom» also means keeping a count of the levels of commitment one has with frequently encountered others. (...) «About what» typically refers to quantifiable first-order resources that the commitment insures access to, or to appropriate tasks for getting these resources. «Until when» means that a commitment is generally attached a precise schedule for its fulfillment."
> (Aubé, 1998, p. 4)

Aubé has developed a general call matrix for basic classes of emotions by combining the approaches of Weiner and Roseman. He places Toda's *social urges* in this matrix.



	VALENCE			
	POSITIVE		NEGATIVE	
AGENCY	actual	anticipated	actual	anticipated
none	JOY (happiness)	HOPE (optimism)	SADNESS (depression)	FEAR (anxiety)
others	GRATITUDE (adoration)		ANGER (contempt)	
self	PRIDE (conceit)		GUILT (shame)	

Fig. 7: Call structure for basic emotions (Aubé, 1998, p.4)



	VALENCE			
	POSITIVE		NEGATIVE	
AGENCY	actual	anticipated	actual	anticipated
none	JOY		GRIEF (SORROW)	FEAR ANXIETY
others	GRATITUDE LOVE		ANGER (JEALOUSY)	
self	DEMONSTRATION (PRIDE) CONFIRMATION		HIDING (SHAME) GUILT	

Fig. 8: Assignment of Todas *urges* to the call structure for basic emotions (Aubé, 1998, p. 5)

Aubé concludes that his modified version of Todas *urges* is consistent with essential theories of motivation and his theory of *emotions-as-commitment-operators*. For him, such a control structure is an essential prerequisite for constructing cooperative adaptive agents that can move independently in a complex social environment.

8.2. Wehrle's Partial Implementation of Toda's Theory

Wehrle has translated the basic elements of Toda's social fungus eater into a concrete computer model (Wehrle, 1994). He used the *Autonomous Agent Modeling Environment* (AAME) as a framework for this. The AAME is specifically designed to explore psychological and cognitive theories of agents in concrete environments. AAME includes an object-oriented simulation language that can be used to model complex microworlds and autonomous systems. The system also includes a number of interactive simulation tools that allow the inspection and manipulation of all objects in the system during execution.

For the concrete implementation of social fungus eaters, some additional assumptions were necessary that are not found in Toda: They keep a certain distance from each other in order to avoid conflicts over food finds or ineffective ore collection. On the other hand, they keep loose contact with each other in order to be able to help each other in an emergency.

Instead of pre-programmed *urges* , Wehrle's model uses a cybernetic control loop in which an agent's energy balance is linked to hedonistic elements.

Overall, the social fungus eater model looks like this:

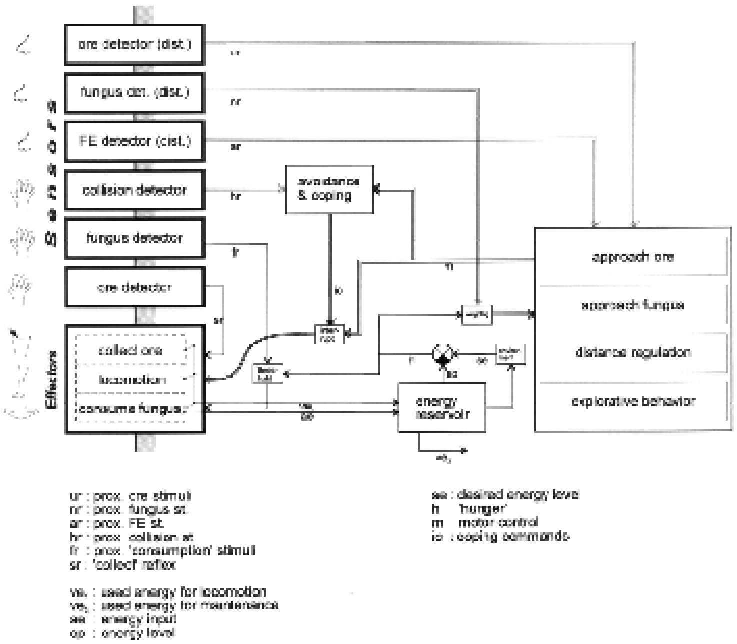


Fig. 9: Model of a social fungus eater (Wehrle, 1994)

Wehrle describes the emergent behavior of the agents in his model as follows:

- Agents usually stay at food sources or ore discovery sites.
- At the food sources, there is a gradual change in agent composition.
- Agents with a similar hunger value form groups of 2 to 5 members.
- The groups disband when the agents have replenished a certain energy level or other agents come to the food source.
- Extremely hungry agents push less hungry agents aside and exhibit other types of antisocial behavior.

Certainly, this implementation is only a small first step towards actually developing an autonomous system according to Toda's principles. However, it shows that it is possible in principle and that the first emergent effects can already be seen with a very restrictive implementation.

8.3. Pfeifers “Fungus-Eater-Principle”

On the basis of the model developed by Toda, Pfeifer has described the construction of autonomous agents according to the fungus-eater principle in a number of papers (Pfeifer, 1994, 1996).

The starting point for him was the only partial success of his model FEELER (see above) as well as other attempts by "classical" AI to design computer models for emotions. In total, he names six points of criticism of such models:

1) The assumption that emotions are isolated phenomena: On the one hand, this provides a number of sources of error, because there is no generally accepted definition of "emotion". On the other hand, "emotional behavior" is firmly programmed into the system via rules; emergent emergence of emotions is therefore not possible. However, there are many indications that emotions are emergent phenomena that cannot be separated from the overall system of an agent.
2) The *frame-of-reference* problem: The models typically used in AI are intentional in nature, i.e. they work with goals, knowledge, convictions. These models tell us nothing about the underlying mechanisms of an emotion, as they are post-hoc rationalizations. Thus, they are attributions emanating from the observer and not images of the emotion mechanisms.
3) Embedding in a real world (*situationality*): The knowledge of how an agent has to react in a certain life situation is not stored once and for all, but is generated again and again in such situations. In an uncertain, rapidly changing, and unpredictable environment, it is not possible to store solutions to all problems in the system from the start.
4) *Embodiment*: Most AI models work exclusively with simulation within the framework of a software model. But lifelike agents have a body and move around in their world with it. The ability to interact with the world through a body generates completely new learning and problem-solving effects that cannot be derived from pure software modeling.
5) Limited tasks: AI models construct their agents for a narrowly defined task. (In the case of FEELER, it was an emotion-inducing situation on an airplane.) This has nothing to do with the real world, where an agent always has to perform multiple tasks, often from different problem areas. A complete autonomous system therefore needs devices to be able to interact with the real world and mechanisms that allow it to act truly autonomously.
6) Overdesign: Apparent complexity in observable behavior does not necessarily mean the same complexity in the underlying design. Braitenberg has already demonstrated this with his *vehicles* (Braitenberg, 1990). Most AI models tend to implement complex rather than simple solutions because they take a top-down approach that starts from hypotheses about a mechanism without letting it develop in the agent itself.

Pfeifer's fungus-eater principle assumes that intelligence and emotions are properties of "complete autonomous systems". That's why he's interested in designing such systems. This also avoids having a fruitless debate about emotions and their function:

> “Since emotions are emergent, they are not engineered into the system, which implies that there can be no set of basic emotions out of which all the others are composed. Identifying the basic components would also imply the existence of clearly delineable functional components which, given that emotions are emergent, is not a sensible assumption.
> Another example concerns the function of emotion. If there is no emotion component we cannot sensibly be talking about its function. What we can say is that the way the complete system is organized enables it to display certain adaptive behaviors. And a convenient way of describing this behavior and communicating about it is in terms of emotion.”
> (Pfeifer, 1994, S. 16)

At the same time, the fungus-eater principle means that you have to observe a correspondingly constructed agent over a longer period of time to see which forms of behavior arise under which conditions.

Based on these explanations, Pfeifer has developed two models of an autonomous Fungus Eater: a *Learning Fungus Eater* with physical implementation and a *Self-sufficient Fungus Eater* as a pure software simulation.

The *Learning Fungus Eater* is a small robot equipped with three types of sensors: *proximity sensors* detect the distance to an obstacle (high activation at proximity, low at distance); *collision* detectors are activated in the event of collisions; *target sensors* can detect a target if they are within a certain radius of the target. The robot has two wheels that are driven independently of each other by two motors.

The *Learning Fungus Eater* has two reflexes: *collision-reverse-turn* and *if-target-detected-turn-towards-center-of-target*. The control architecture consists of a neural network, which can be modified in part by Hebb's learning:

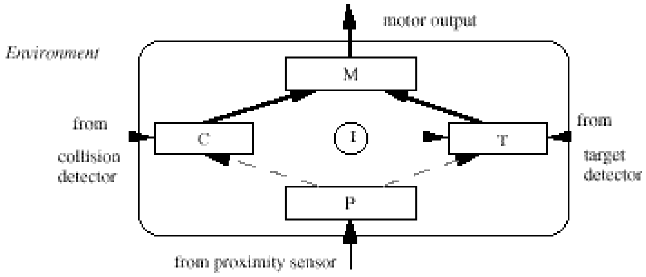


Fig. 10: Control architecture of the *Learning Fungus Eater* (Pfeifer, 1994, p. 10)

The entire control system consists of four layers: the *collision layer*, the *proximity layer*, the *target layer* and the *motor layer*. The robot's only task is to move. Its environment looks like this:

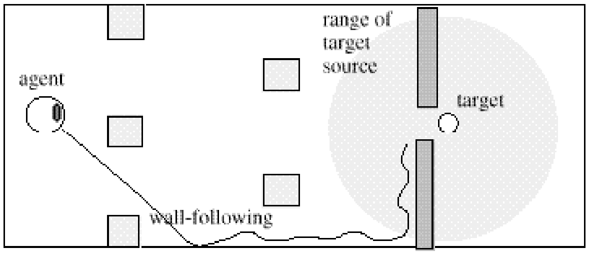


Abb. 11: Environment of the *Learning Fungus Eater* (Pfeifer, 1994, S.10)

What happens when the robot is activated? First, he will crash into obstacles. Each *reverse-and-turn* action makes Hebbian learning possible between the *proximity layer* and the *collision layer* until the robot has learned to avoid obstacles. From the outside, it looks as if the robot can anticipate obstacles. Pfeifer points out that an architecture with several layers is usually proposed for such "anticipating" behavior, while in fact one layer is sufficient. Pfeifer explains two other phenomena of the robot:

> “What also happens if the “Fungus Eater” moves about is that if there are target sources (light) along walls, it will start following walls, even in the absence of light. We can characterize the behavior adopting the intentional and the “emotional stance”. People who observed the robot’s behavior have made remarks like the following (sic!): “Oh, it’s following walls because it knows that food is normally along walls, so it’s a good strategy to follow walls.” “It hopes to find food because it has learned that food is normally found along walls.” If the “Fungus Eater” is dragged into a corner with a light source where it can no longer move it will wiggle back and forth in the corner for some time and then turn around and take off. People have said that it was frustrated or annoyed and turned away.”
> (Pfeifer, 1994, S. 11)

This clearly shows, according to Pfeifer, that a system without strategies, without an anticipation mechanism, without knowledge of food sources can show behavior that is classified by observers as goal-oriented and motivated - but which results only from the interaction of the system with its environment.

The *Learning Fungus Eater* is not a complete autonomous system in that it cannot be self-sufficient. That's why Pfeifer has developed the *Self-sufficient Fungus Eater* , but initially only as a software simulation.

In this case, the agent is in a Toda landscape with mushrooms for food and ore to mine. The choice of action is much more complicated here: the agent can explore (search for ore or food), he can eat or collect ore. What it does is determined by the central variables "energy level" and "amount of ore collected". For the choice of action in any given situation, the agent has only one rule: "If the agent is exploring and energy level is higher than amount of ore collected per unit time, it should ignore fungus (but should not ignore ore), if not it should ignore ore (but should not ignore fungus)." (Pfeifer, 1994, p. 12)

Here, too, according to Pfeifer, the result is a state of affairs for observers to which they ascribe a high emotional intensity:

> “If they see, e.g. energy level going down.... and they see the agent moving toward a patch of fungus.... they really get excited about whether it will make it or not. Such situations are normally associated by observers with emotion: there is time pressure for the agent which may be associated with fear or at least with an increasing level of agitation (This is a typical consequence of self-sufficiency). However, we know that all there is in terms of mechanism within the agent is the simple rule. Thus, if we want to talk about emotion at all, it is emergent.... In spite of its simplicity the “Self-sufficient Fungus Eater” shows in some situations behavior that we might associate with high emotional intensity.”
> (Pfeifer, 1994, S. 13)

In the same essay, Pfeifer qualifies that these few findings cannot, of course, explain what emotions really are. However, he expects more from the continuation of this approach, even if it is very time-consuming, than from computer models that have an isolated emotion model as their subject.

8.4. The approach of Dörner et al.

Dörner has developed a computer model that integrates cognitive, motivational and emotional processes (Dörner et al., 1997; Dörner and Schaub, 1998; Schaub, 1995, 1996): the PSI model of intention regulation. Within the framework of PSI, the "EmoRegul" model has been developed, which takes special account of emotional processes.

PSI is part of a theoretical approach that Dörner calls "synthetic psychology". This approach attempts to analyze what mental processes look like as processes of information processing by constructing these psychic processes. Dörner's starting point is similar to Toda's when he writes, ".. that in psychology one must not dissect the various psychological processes into their components with impunity" (Dörner and Schaub, 1998, p.1).

At the core of the PSI model is the concept of "intention". Schaub defines intention as an internal psychological process,

> "... which is defined as an ephemeral structure consisting of an indicator of a state of deficiency (hunger, thirst, etc.) and processes for eliminating or avoiding this state of deficiency (either ways to *consumptive final action* or ways to avoid, in the broadest sense "*flight*").
> (Schaub, 1995, p. 1)

The PSI agents are designed as steam engines that move in a simulated environment with water points, petrol stations, quarries, etc. In order to continue moving, the agents need both gasoline and water, which can be found in different places.

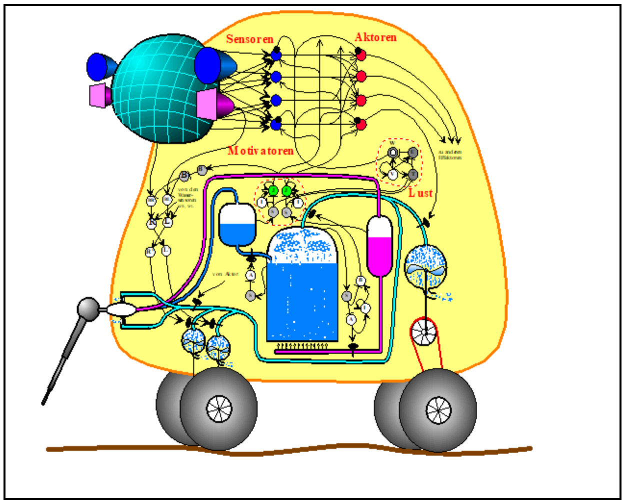


Fig. 12: Schematic structure of the PSI system (Dörner and Schaub, 1998, p. 7)

A psi agent has a set of needs that are divided into two classes: material needs and informational needs. Material needs include the intake of fuel and water as well as the avoidance of dangerous situations (e.g. falling rocks). Informational needs include determinacy (an expectation is fulfilled) and indeterminacy (an expectation is disappointed), competence (fulfilment of needs) and affiliation (need for social contacts).

In PSI, deficiencies can be imagined as a vessel whose contents have fallen below a certain threshold value. Dörner refers to the difference between the actual and target state as demand. "A need therefore signals that there is a need of a certain extent." (Dörner and Schaub, 1998, p. 10)

Such a deficiency state activates a motivator, whose degree of activation is the higher the greater the setpoint deviation and the longer it has lasted. The motivator now tries to take over the control of action in order to eliminate this state through a consumptive final action. To this end, a goal is pursued that is known to the motivator from past experience.

Goals in PSI also create motifs that represent instances that set action in motion, align it with a specific goal and maintain it until the goal is achieved.

Since several motivators are always vying with each other for the direction of action, the system has a motive selector that decides with the help of a quick expectation x value calculation which motivator has the greatest motive strength and should thus be given preference. The value of a motivator is determined by its importance (size of the setpoint deviation) and urgency (available time until the target state is eliminated); the expectation is determined by the agent's ability to actually satisfy this need (probability of success).

PSI also has a memory that consists of sensory and motor "schemes". Sensory schemes represent the agent's knowledge of his environment; motor schemata are behavioral programs. In PSI, there is no distinction between different types of memory.

Action control in PSI takes place via intentions, which are defined operationally as a combination of the selected motif with the information linked to the active motivator in the memory network. This information concerns the goals to be pursued, the operators or action schemes to be used, the knowledge of the past, futile approaches to problem solving, and the plans that PSI generates with the help of heuristic methods. All this information consists of neural networks; an intention as a bundling of all this information is the working memory of PSI.

The central mechanisms of emotional regulation in psi are the motivators for determination and competence, i.e. two informational needs. Active determination or competence motivators trigger certain actions or increase the willingness to do so:

> "A decrease in specificity leads to an increase in the extent of "background control". This means that PSI turns away from its current intention more often than usual and controls the environment. Because with little predictability of the environment, you should be prepared for anything. (...) Furthermore, with decreasing determination, the tendency to escape behaviors or to behaviors of specific exploration increases. (...) Not so extreme cases of flight are refusal to provide information; one simply no longer looks at the areas of reality that have proven to be indeterminate. The "withdrawal" from reality also includes the fact that PSI becomes more hesitant in its behavior when its determination decreases, does not move so quickly to action, plans longer than it would under other circumstances, is not so "courageous" in exploring."
> (Dörner and Schaub, 1998, p. 33)

In PSI, emotions do not arise in a separate emotion module, but as a result of control processes of a homeostatic system. Schaub puts it this way: "What we call emotions in humans is the way in which action is organized, combined with associated motivations." (Schaub, 1995, p. 6)

Dörner admits that a variety of emotions cannot yet be represented in PSI because the system lacks a module for self-observation and self-reflection. However, this is only a question of the refined implementation of the model and therefore not a problem in principle.

Dörner's model has a number of similarities with other models. Like Toda and Pfeifer, his starting point is to construct a complete autonomous system without a separate emotion module. Like Frijda and Moffat, PSI contains a central memory that is accessible to all modules at any time for reading and modification:

> "The use of common memory structures allows all processes to obtain information, especially about the intentions to be processed. Each sub-process is thus aware, for example, of the importance and urgency of the current intention."
> (Schaub, 1995, p. 6)

PSI does not work with explicit rules, but is a partially connectionist system that produces emotions through self-organization.

8.5. Summary and evaluation

The model approaches of Pfeifer and Wehrle clearly show the significance of Toda's theory for the construction of autonomous agents that could not survive without the control function of emotions. Instead of predefined emotion taxonomies that are firmly anchored in the model, both authors take the opposite approach: their models contain only the most necessary instructions for the agent.

While Wehrle still links certain events to hedonistic components, Pfeifer does not do so at all. As a result, both systems show behavior that can be interpreted as "emotional" by an outsider.

Both models have the disadvantage that they do not say too much about emotions in computer agents - this requires a longer observation period in which the agents can develop. This avoids the problem of arbitrarily programming emotions into a system; on the other hand, a new front of discussion is opened about whether a behavior that appears to an observer to be emotional is actually emotional. Here the argumentation moves clearly into the philosophical realm again.

Both approaches consistently think the assumption of emotions as emergent phenomena to the end - with all the advantages and disadvantages that result from this.

Aubé's attempt to link Todas *urges* to his theoretical model of emotion has its own problems. He correctly recognizes a number of inconsistencies in Toda's *urges* model and tries to eliminate them. In doing so, however, he puts his own definition of emotions as social phenomena in the foreground. Aubé's fusion of Weiner's and Roseman's theories, which he then summarizes again with his and Toda's approach to form a unity, raises fundamental problems for which there is no space at this point.

Finally, Dörner's model is similar in many respects to the approaches of Pfeifer and Wehrle (and thus Toda): emotions are understood as control functions in an autonomous system. Dörner combines this approach with a homeostatic regulation model. Dörner also does not explicitly define emotions; emotional behavior arises due to the change of two parameters, which he calls "determination" and "competence". In this respect, emotional action is only attributed to the system from the outside. Dörner also considers emotions as emergent phenomena that do not have to be integrated into a system as a separate module. It remains to be seen in which direction PSI (and thus Dörner's emotion model) will develop when the system receives a self-observation module.

9. The Philosopher from Birmingham

Aaron Sloman, professor of philosophy at the University of Birmingham's School of Computer Science, is certainly one of the most influential theorists on computational models of emotions. As early as 1981, he announced in the title of an essay: "Why robots will have emotions". His reasoning was:

> “Emotions involve complex processes produced by interactions between motives, beliefs, percepts, etc. E.g. real or imagined fulfilment or violation of a motive, or triggering of a ’motive-generator’, can disturb processes produced by other motives. To understand emotions, therefore, we need to understand motives and the types of processes they can produce. This leads to a study of the global architecture of a mind.”
> (Sloman, 1981, p.1)

Like Bates, Reilly or Elliott, Sloman also advocates the *broad and shallow* approach. For him, it is more important to develop a complete system with little depth than individual modules with a lot of depth. In his opinion, this is the only way to create a model that reflects reality in a reasonably realistic way.

Since 1981, Sloman and his collaborators in the *Cognition and Affect Project* have published a large number of papers on the topic of "intelligent systems with emotions", which can be broadly divided into three categories:

1. Work that deals with the basic approach to the construction of an intelligent system;
2. work that deals with the basic elements of such a system and
3. Works that try to implement such a system in practice.

In order to be able to classify Sloman's remarks correctly, one must place them in the context of his (epistemological) theoretical approach, which does not primarily revolve around emotions, but around the construction of intelligent systems.

In the following, we will try to briefly present the core ideas of Sloman's theory, as they form the basis for understanding Ian Wright's "libidinal computer" (see below).

9.1. Approaches to the design of intelligent systems

Sloman's interest in his work is not primarily in a simulation of the human mind, but in the development of a general "intelligent system", independent of its physical substance. Humans, bonobos, computers and aliens are different implementations of such intelligent systems - but the underlying construction principles are identical.

Sloman divides previous attempts to develop a theory about the workings of the human mind (and thus intelligent systems in general) into three broad groups: semantics-based, phenomena-based and *design-based*.

Semantics-based approaches analyze how people describe mental states and processes in order to determine implicit meanings that underlie the use of everyday language words. Among them he counts the approaches of Ortony, Clore and Collins as well as Johnson-Laird and Oatley. Sloman's argument against these approaches is: "As a source of information about mental processes such enquiries restrict us to current 'common sense´ with all its errors and limitations." (Sloman, 1993, p. 3)

According to Sloman, some philosophers who analytically examine concepts also produce semantics-based theories. What distinguishes them from psychologists, however, is the fact that they do not concentrate solely on *existing* concepts, but are often more interested in the set of all *possible* concepts.

Phenomenon-based approaches assume that psychic phenomena such as "emotion", "motivation" or "consciousness" are already clear and that everyone can intuitively recognize concrete examples of them. They therefore only try to correlate simultaneously occurring and measurable phenomena (e.g. physiological effects, behaviour, environmental characteristics) with the occurrence of such psychological phenomena. These approaches, according to Sloman, are particularly found among psychologists. His criticism of such approaches is:

> “Phenomena-based theories that appear to be concerned with mechanisms, because they relate behaviour to neurophysiological structures or processes, often turn out on close examination to be concerned only with empirical correlations between behaviour and internal processes: they do not show why or how the mechanisms identified produce their alleged effects. That requires something analogous to a mathematical proof, or logical deduction, and most cognitive theories fall far short of that.”
> (Sloman, 1993, p. 3)

Design-based approaches break the boundaries of these two approaches. Sloman explicitly refers here to the work of the philosopher Daniel Dennett, who has shaped the debate about intelligent systems and consciousness like no other.

Dennett distinguishes between three approaches to making predictions about an entity: *physical stance*, *design stance* and *intentional stance*. The *physical stance* is "simply the standard laborious method of the physical sciences" (Dennett, 1996, p. 28); the *design stance*, on the other hand, assumes "that an entity *is* designed as I suppose it to be, and that it will operate according to that design" (Dennett, 1996, p. 29). Finally, *the intentional stance*, which according to Dennett can also be regarded as a "subspecies" of *the design stance*, predicts the behavior of an entity, for example a computer program, "*as if it* were a rational agent" (Dennett, 1996, p. 31).

Proponents of the design-based approach start from the position of an engineer trying to construct a system that produces the phenomena to be explained. However, not every design requires a designer at the same time:

> “The concept of “design" used here is very general, and does not imply the existence of a designer. Neither does it require that where there's no designer there must have been something like an evolutionary selection process. We are not primarily concerned with origins, but with what underlies and explains capabilities of a working system.”
> (Sloman, 1993, p.4)

A design is, strictly speaking, nothing more than an abstraction that determines a class of possible instances. Nor does it necessarily have to be implemented concretely or materially - although its instances may well have a physical form.

For Sloman, the concept of design is closely linked to the concept of niche. A niche is also not a material entity and not a geographical region. Sloman defines them in a broad sense as a collection of requirements for a functioning system.

With regard to the development of intelligent agents in AI, design and niche play a special role. Sloman speaks of *design-space* and *niche-space*. A lifelike intelligent system will interact with its environment and change in the course of its evolution. This allows it to move on a certain *trajectory* through the *design-space*. This also corresponds to a certain trajectory through *niche-space*, because the changes in the system allow it to occupy new niches:

> “A design-based theory locates human mechanisms within a space of possible designs, covering both actual and possible organisms and also possible non-biological intelligent systems.”
> (Sloman, 1991, p. 5)

Sloman identifies different trajectories through the *design-space*: individuals who can adapt or change go through so-called *i-trajectories*. He refers to evolutionary developments that are only possible over generations of individuals as *e-trajectories*. And finally, there are changes to individuals that are made from the outside (for example, debugging software), which he calls *r-trajectories* (r for repair).

Taken together, these elements result in dynamic systems that can be implemented in different ways.

> “Since niches and designs interact dynamically, we can regard them as parts of virtual machines in the biosphere consisting of a host of control mechanisms, feedback loops, and information structures (including gene pools). All of these are ultimately implemented in, and supervenient on physics and chemistry. But they and their causal interactions may be as real as poverty and crime and their interactions.”
> (Sloman, 1998b, S. 6)

For Sloman, one of the most urgent tasks is to clarify biological terms such as niche, genotype, etc., in order to be able to understand the relationships between niches and designs for organisms. This is also a significant advance for psychology:

> “This could lead to advances in comparative psychology. Understanding the precise variety of types of functional architectures in design space and the virtual machine processes they support, will enable us to describe and compare in far greater depth the capabilities of various animals. We’ll also have a conceptual framework for saying precisely which subsets of human mental capabilities they have and which they lack. Likewise the discussion of mental capabilities of various sorts of machines could be put on a firmer scientific basis, with less scope for prejudice to determine which descriptions to use. E.g. instead of arguing about which animals, which machines, and which brain damaged humans have consciousness, we can determine precisely which sorts of consciousness they actually have.”
> (Sloman, 1998b, S. 10f.)

Sloman admits that the requirements for design-based approaches are not trivial. He lists five requirements that such an approach should meet:

1. Analysis of the requirements for an autonomous intelligent agent;
2. a design specification for a functioning system that meets the requirements of (1);
3. a detailed implementation or specification for such an implementation of a functioning system;
4. a theoretical analysis of the extent to which the design specification and the details of the implementation meet or do not meet the requirements, and
5. an analysis of the neighbourhood in *design-space*.

A design-based approach does not necessarily have to be a top-down approach. Sloman expects the most success from models that combine *top-down* and *bottom-up*.

For Sloman, design-based theories are more effective than others because:

> "Considering alternative possible designs leads to deeper theories, partly because the contrast between different design options helps us understand the trade-offs addressed by any one design, and partly because an adequate design-based theory of human affective states would describe mechanisms capable of generating a wide range of phenomena, thereby satisfying one of the criteria for a good scientific theory: generality. Such a theory can also demonstrate the possibility of new kinds of phenomena, which might be produced by special training, new social conditions, brain damage, mental disturbance, etc."
> (Sloman, 1991, p. 5)

9.2. The basic architecture of an intelligent system

What a design-based approach creates are architectures. Such an architecture explains which states and processes are possible for a system that possesses this architecture.

Aus der Menge aller möglichen Architekturen ist Sloman an einer bestimmten Klasse besonders interessiert: "..."high level" architectures which can provide a systematic non-behavioural conceptual framework for mentality (including emotional states)." (Sloman, 1998a, S. 1) Ein solches *framework for mentality*

> "is primarily concerned with an "information level" architecture, close to the requirements specified by software engineers. This extends Dennett's "design stance" by using a level of description between physical levels (including physical design levels) and "holistic" intentional descriptions."
> (Sloman, 1998a, S. 1)

According to Sloman, an architecture for an intelligent system consists of four essential components: several functionally different processing layers, control states, motivators and filters, and a global alarm system.

9.2.1. The processing layers

Sloman assumes that every intelligent system has three processing layers:

- a reactive layer that includes automatic and hard-wired processes;
- a deliberative layer for planning, evaluating, allocating resources, etc.;
- a meta-management layer that includes observation and evaluation mechanisms for internal states.

The reactive layer is the oldest in evolutionary history, and there are a large number of organisms that only have this layer. Schematically, a purely reactive agent looks like this:

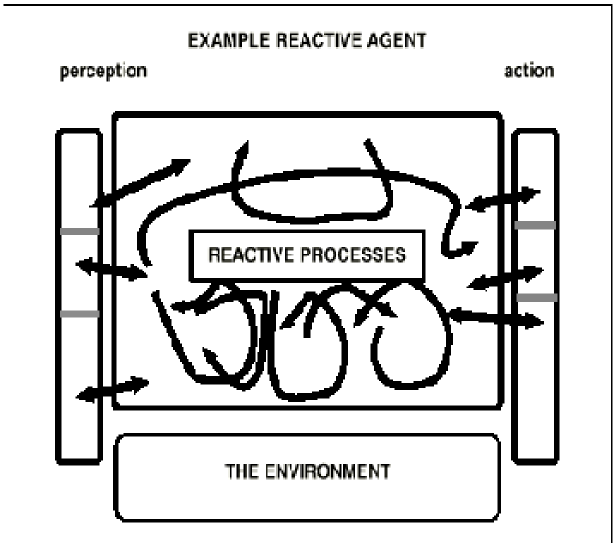


Fig. 13: Reactive architecture (Sloman, 1997a, p. 5)

A reactive agent cannot make plans or develop new structures. It is optimized for very special tasks; however, he cannot cope with new tasks. What it lacks in flexibility, however, it gains in speed. Since almost all processes are clearly defined, its reaction speed is high. Insects, according to Sloman, are an example of such purely reactive systems, which at the same time prove that the interaction of numerous such agents can produce amazingly complex results (e.g. termite towers).

A second, phylogenetically younger layer gives an agent far more qualities. Schematically, it looks like this:

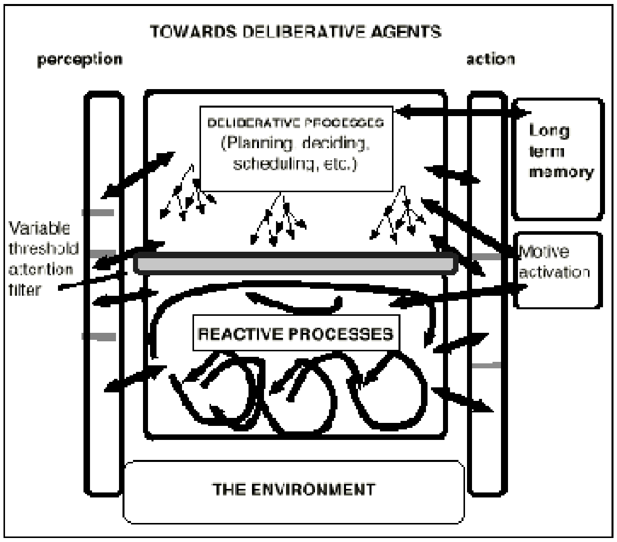


Abb. 14: Deliberative architecture (Sloman, 1997a, S. 6)

A deliberative agent can recombine his repertoire of actions at will, develop plans and evaluate them before execution. An essential prerequisite for this is a long-term memory in order to store plans that have not yet been completed or to store the probable consequences of plans and evaluate them later.

The construction of such plans proceeds step by step and is therefore not a continuous, but a discrete process. Many of the processes in the deliberative layer are serial in nature and thus resource-limited. This seriality offers a number of advantages: the system is always aware of which plan has led to success and can allocate rewards accordingly; the execution of contradictory plans at the same time is prevented; communication with the long-term memory is largely error-free.

Such a resource-limited subsystem is of course highly susceptible to failure. Therefore, filtering processes with variable thresholds are necessary to ensure that the system works (see below).

The phylogenetically youngest layer of the overall system is what Sloman calls meta-management:

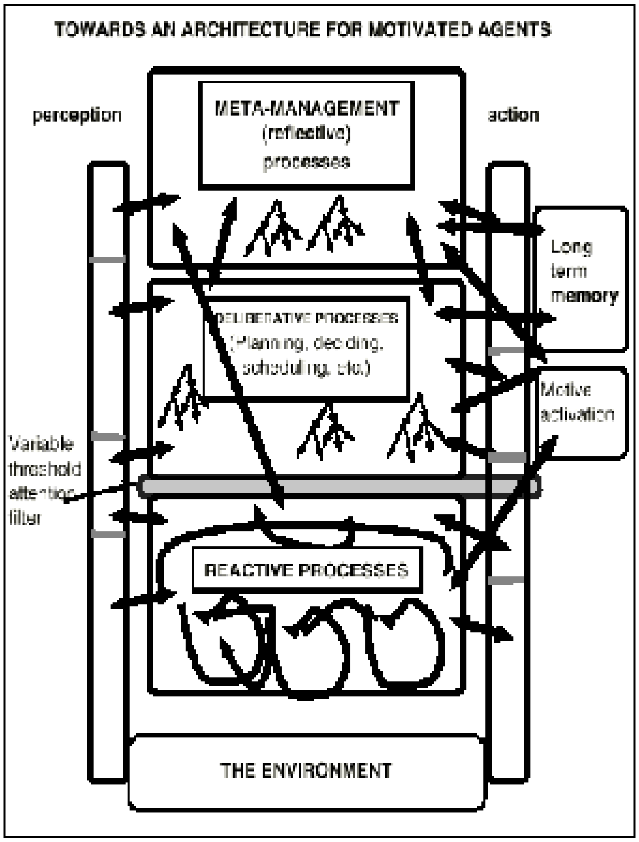

TOWARDS AN ARCHITECTURE FOR MOTIVATED AGENTS
perception
action
META-MANAGEMENT (reflective) processes
Long term memory
DELIBERATIVE PROCESSES (Planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
REACTIVE PROCESSES
THE ENVIRONMENT

Abb. 15: Meta-Management Architektur (Sloman, 1997a, S. 7)

This is a mechanism that tracks and evaluates the internal processes of the system. Such a subsystem is necessary to evaluate and, if necessary, discard the plans and strategies developed in the deliberative layer, to recognize recurring patterns in the deliberative subsystem, to develop long-term strategies, and to communicate effectively with others.

Sloman points out that these three layers are not hierarchically structured, but parallel and that they also work in parallel. Like the overall system, these modules also have their own architecture, which can include other subsystems with their own architecture.

The Meta Management module is anything but perfect in its work. This is because it does not have comprehensive access to all internal states and processes, that control over the deliberative subsystem is incomplete, and that self-evaluations can be based on false premises.

9.2.2. The Control States

An architecture such as the one outlined so far has a variety of control states at different levels. Some of them operate at the highest level of abstraction, while others take place unconsciously in frequent control decisions.

The following figure gives an overview of the control states of the system:

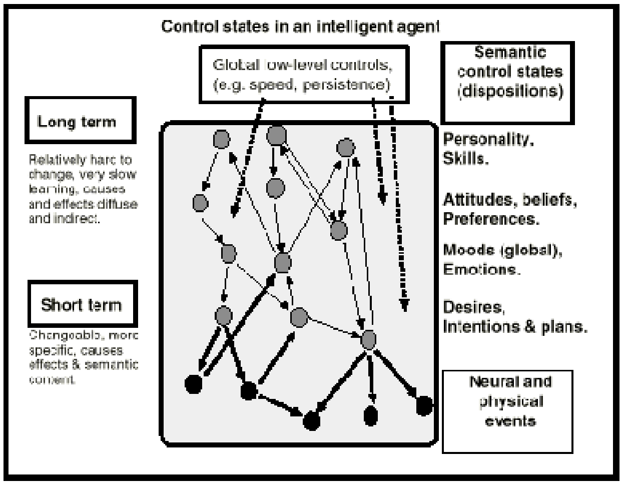

Control states in an intelligent agent
Global low-level controls, (e.g. speed, persistence)
Semantic control states (dispositions)
Long term
Relatively hard to change, very slow learning, causes and effects diffuse and indirect.
Short term
Changeable, more specific, causes effects & semantic content.
Personality, Skills.
Attitudes, beliefs, Preferences.
Moods (global), Emotions.
Desires, Intentions & plans.
Neural and physical events

Fig. 16: Control states of an intelligent system (Sloman, 1998b, p. 17)

Different states of control also have different underlying mechanisms. Some can be chemical in nature, while others have to do with information structures.

Control states contain dispositions to react to internal or external stimuli with internal or external actions. Within the framework of the overall system, numerous control states can exist simultaneously and interact with each other.

States of control are known by *numerous names in folk psychology*: desires, preferences, beliefs, intentions, moods, etc. By defining such states through architecture, Sloman wants to provide a "rational reconstruction of many everyday mental concepts".

Each control state has, among other things, a structure, transformation possibilities and, if necessary, semantics. Sloman illustrates this with the example of a motivator (see below):

> “For example, a motivator may have a complex internal structure (syntax), a content (based on that structure) referring to certain states of affairs (semantics), a functional role, i.e. dispositional powers to determine internal and external actions (pragmatics). It may also enter into processes that derive new motivators or plans (inference), it may be brought about or triggered in various ways (aetiology), and may be modified, suppressed or terminated by other processes (liabilities). Some control states are short-lived (e.g., a motivator which is immediately rejected, or whose goal is quickly achieved). Others endure.”
> (Wright et al., 1996, S. 12)

Control states also differ in whether they are easy or very difficult to change. According to Sloman, many higher-order control states can only be modified in small steps and over a longer period of time. Higher-order control states are also more powerful and influential with regard to the overall system than lower-order control states.

Sloman postulates a process called *circulation*, through which the control states circulate through the overall system. Useful states of control can rise up the hierarchy and increase their influence; useless states of control can almost completely disappear from the system.

> “Control states may be qualitatively transformed during circulation, for instance acquiring more general conditions of applicability. Higher level general attitudes such as generosity of spirit, may also spawn derivative specialised control states such as favouring a certain political party - another aspect of circulation. Internal connections between control states will set up suppressive or supportive relationships, dependencies, mutual dependencies and, occasionally, dead-locks.”
> (Wright et al., 1996, S. 12)

The result of all these processes is a kind of diffusion, with which the effects of a strong motivator are slowly distributed into countless and long-lasting *control sub-states* , up to irreversible integration into reflexes and automatic reactions.

9.2.3. Motivators and Filters

Motivators are a central component of any intelligent system. Sloman defines them as "mechanisms and representations that tend to produce or modify or select between actions, in the light of beliefs." (Sloman, 1987, p. 4).

Motivators can only arise if goals are present. A goal is a symbolic structure (not necessarily physical in nature) that describes a state that is to be achieved, maintained, or prevented. While beliefs are defined by the fact that they are representations that adapt to reality through processes of perception and deliberation, goals are representations that trigger behavior in order to adapt reality to representation.

Motivators are generated by a mechanism that Sloman calls *motivator generator* or *motivator generactivator*. Motivators are generated on the basis of external or internal information or produced by other motivators. Sloman formally defines a motivator structure in terms of ten fields: (1) a possible state that can be true or false; (2) a motivational attitude towards this condition; (3) an assumption (*amount)* about this condition; (4) an *importance value*; (5) an urgency; (6) an insistence value; (7) one or more plans; (8) a *commitment status*; (9) management information and (10) a dynamic state such as "plan postponed" or "currently under consideration".

In a later work (Sloman, 1997c), Sloman extended this structure by two further fields: (11) a rational reason, if the motivator has arisen from an explicit thought process, and (12) an intensity that determines whether a motivator that has already been worked on continues to be preferred over other motivators.

The strength of a motivator is determined by four variables:

a. The insistence determines the probability with which the motivator can overcome the filter (see below);
b. importance determines the likelihood that the motivator will subsequently be accepted and persecuted;
c. the intensity determines how actively and intensively a motivator is pursued if he is accepted;
d. urgency determines the point in time up to which the motivator must have been persecuted.

Motivators compete with each other for attention, i.e. for the limited resources of the deliberative subsystem. For this subsystem to work, there must be a mechanism that prevents new motivators from attracting attention at any time. For this purpose, the system has a so-called *variable threshold attention filter*.

The filter sets a threshold value that a motivator must overcome in order to be able to claim attention resources for himself. As the name implies, this filter is variable and can change, for example through learning. Sloman explains this with the example of a novice driver who cannot talk to anyone else while driving because he has to concentrate too much on the road. However, after a certain amount of practice, he is able to do so.

The insistence of a motivator, i.e. the decisive variable for overcoming the filter, is a quickly calculated heuristic value of the importance and urgency of the motivator.

When a motivator has *surfaced*, i.e. has overcome the filter, several management processes become active. Such management is necessary because several motivators always pass through the filtering process. These processes are *adoption-assessment* (the decision whether to accept or reject a motivator); *scheduling* (deciding when to execute a plan for that motivator); *expansion* (developing plans for the motivator) and *meta-management* (deciding whether and when a motivator should be considered by management at all).

Sloman's *attention filter penetration theory* requires a higher degree of complexity than the theory of Oatley and Johnson-Laird. He postulates that not every motivator interrupts the current activity, but only those that either have a high degree of *insistence* or are not particularly high *for the corresponding* attention filters.

> “High insistence of a new motive can cause attention to be diverted without actually causing any current action to be interrupted or disturbed. For example feeling very hungry can make a driver consider whether to stop for a meal, without interfering with the driving. Interruption might occur if the new goal is judged more important than, and inconsistent with, the purpose of the current activity, or if the new one is judged to be very urgent (although not necessarily very important) whereas the (more important) current activity is not time-critical and can be temporarily suspended: for instance stopping for a meal because one has plenty of time before the important meeting. Alternatively a highly insistent motive that gets through the filters can be considered and then rejected as relatively unimportant, without interrupting any important current action. So insistence, the propensity to divert attention, is not the same as a propensity to interrupt current actions, except those that require full attention.”
> (Sloman, 1992c, S. 15)

9.2.4. The Global Alert System

A system that has to survive in an environment that is constantly changing needs a mechanism by which it can react to such changes without much time delay. Such a mechanism is an alarm system.

An alarm system is important not only for a reactive, but also for a deliberative architecture. For example, planning ahead can highlight a danger or an opportunity to which a change of strategy must be responded to immediately.

Sloman draws a parallel between his alarm system and neurophysiological findings:

> "Our global alarm mechanism corresponds closely to the assumed role of the limbic system including the amygdala which is thought to learn associations of the type involved in emotions."
> (Sloman, 1998e, S.4)

The different layers of the system are affected by the alarm system, but in different ways. At the same time, they can also transfer information to the alarm system themselves and thus trigger a global alarm.

9.3. Emotions

For Sloman, emotions are not independent processes, but arise as an emergent phenomenon from the interaction of the different subsystems of an intelligent system. Therefore, there is no need for an independent "emotion module".

A consideration of psychological theories of emotion leads Sloman to the conclusion:

> "Disagreements about the nature of emotions can arise from failure to see how different concepts of emotionality depend on different architectural features, not all shared by all the animals studied."
> (Sloman and Logan, 1998, S. 6)

If, on the other hand, emotions are considered as the result of appropriately designed architecture, many misunderstandings can be cleared up, according to Sloman. For him, a theory that analyzes emotions in connection with architectural concepts is therefore more effective than other approaches:

> "This, admittedly still sketchy, architecture, explains how much argumentation about emotions is at cross-purposes, because people unwittingly refer to different sorts of mechanisms which are not mutually exclusive. An architecture-based set of concepts can be made far less ambiguous."
> (Sloman and Logan, 1998, p. 7)

The different layers of the sketched architecture also support different emotions. The reactive layer is responsible for disgust, sexual arousal, startle and frightening of large, rapidly approaching objects. The deliberative layer is responsible for frustration through failure, relief through danger avoidance, fear of failure, or pleasant surprise through success. The meta-management layer supports shame, humiliation, aspects of grief, pride, anger.

In doing so, Sloman's approach deliberately ignores the physical side effects of emotions. For him, these are only peripheral phenomena:

> "They are peripheral because essentially similar emotional states, with similar social implications, could occur in alien organisms or machines lacking anything like our expression mechanisms."
> (Sloman, 1992c, S. 20)

Sloman also does not want to accept the objection that emotions are inseparable from physical expression. He counters with the argument that these are "relics of our evolutionary history" that have no fundamental significance for emotions. An emotion does not derive its meaning from the physical feelings that accompany it, but from its cognitive content:

> "Fury matters because it can produce actions causing harm to the hater and hated, not because there is physical tension and sweating. Grief matters because the beloved child is lost, not because there's a new feeling in the belly."
> (Sloman, 1987, p. 9)

He argues similarly with regard to a number of non-cognitive factors that can play a role in human emotions, for example chemical or hormonal processes. He asks whether the affective states triggered by such non-cognitive mechanisms are really so different from those produced by cognitive processes:

> "Is a mood of depression or euphoria that is produced by chemical processes a totally different state from the depression produced by repeatedly failing to pass your examinations or the euphoria produced by passing with distinction?"
> (Sloman, 1992c, S. 21)

So how do emotions arise in Sloman's intelligent system? Basically, he distinguishes between three classes of emotions that correspond to the three layers of his system. On the one hand, emotions can arise from internal processes within one of these layers; on the other hand, through interactions between the layers.

Emotions are often accompanied by a state that Sloman calls "perturbanz" (*perturbance*). A perturbanz is given when the overall system is partially out of control. It always occurs when a rejected, postponed or simply unwanted motivator repeatedly appears, thus preventing or complicating the management of other, more important goals.

The decisive factor for this is the insistence value of a motivator, which for Sloman represents a dispositional state. As such, a highly insistent motivator can trigger perturbanzes even if he has not yet overcome the filter or is not yet actively being worked on.

> "Insistence, on this analysis, is a dispositional state: the highly insistent motive or thought need not actually get through the filter and interrupt anything. Even if it does get through it need not actually disturb any current activity. I suggest it is this strong potential for such disturbance and diversion of attention that characterizes many of the states we describe as emotions. Such states can exist whether or not attention is actually diverted, and whether or not actions are thereby interrupted or disturbed. Thus, like jealousy, anger (in the form of a very insistent desire to harm someone because of something he is believed to have done that is strongly negatively evaluated) can persist even though something else occupies attention for

> a while. During that time there is no diversion of attention or disturbance of any action. Dormant dispositions include such emotional states.”
> (Sloman, 1992c, S. 16)

Perturbances can be *occurrent* (an attempt to gain control of attention) or *dispositional* (not an attempt to gain control of attention).

Perturbanted states differ according to several dimensions: duration, internal or external source, semantic content, type of disorder, effect on attention processes, frequency of the disorder, positive or negative evaluation, development of the condition, subsidence of the condition, etc.

Perturbances, like emotions, are emergent effects of mechanisms whose task is to do something else. They are created by the interaction of

- resource-limited, attentive processing;
- a subsystem that produces new candidates for such processing;
- a heuristic filter mechanism.

For the appearance of perturbances, therefore, there is no need for a separate "perturbanz mechanism" in the system; the question of the function of a perturbaned state is also not meaningful from this point of view. However, perturbances are not to be equated with emotions; rather, they are typical side effects of states that are commonly referred to as emotional.

For Sloman, emotional states are basically nothing more than motivational states caused by motivators.

> “Since insistence, as I have defined it, is a matter of degree, the theory implies that there are only differences of degree between emotional and non-emotional motivational states. It also implies that there is much in common between emotional states and those cognitive states where a particular thought or something like a remembered experience or tune has high insistence, but does not involve any particular motivation or positive or negative evaluation.”
> (Sloman, 1992c, S. 17)

Another characteristic of emotional states is that they produce new motivators. For example, if a first emotional state has arisen from a conflict between a belief and a motivator, new motivators can emerge from it, leading to new conflicts within the system.

9.4. The Implementation of the Theory in MINDER1

Sloman and his research group have developed a working computer model called MINDER1, in which his architecture is partially implemented. MINDER1 is a pure software implementation; so there is no crèche with actual robots. The system is described below as an overview; a detailed account can be found in [Wright and Sloman, 1996].

MINDER1 consists of a kind of virtual crèche in which a virtual nanny (the Minder) has to take care of a number of virtual babies. These babies are "reactive minibots" that are constantly wandering around the crèche and are exposed to different dangers: they can fall into ditches and be damaged or die; their batteries can run dry, so they have to go to a charging station; if the batteries are too depleted, they also die; Overcrowding of the crèche turns some babies into bullies, who then harm other babies; damaged babies must be taken to the infirmary for repair; if the damage is too great, the baby dies.

The minder now has different tasks: he has to make sure that the babies do not lose energy, that they do not fall into a ditch or are exposed to other dangers. For this purpose, he can, for example, build fences to lock the babies in. He has to bring minibots that are about to run out of energy to a charging point or others as far away from a ditch as possible.

This variety of tasks ensures that the minder must constantly create, evaluate and act on new motives. The more minibots come into the crèche, the more the performance of the minder decreases.

The architecture of MINDER1 is in line with the basic principles outlined above. It consists of three subsystems, which in turn have a number of other subsystems.

9.4.1. The Reactive Subsystem

The reactive subsystem comprises four modules: *perception*, *belief maintenance*, reactive plan execution and pre-attentive *motive generation*.

The perception subsystem consists of a database that contains only partial information about the environment of the minder. The system works within a certain radius around the Minder, but cannot discover hidden objects, for example. An update of the database looks like this:

```
[new_sense_datum
time 64 name minibot4 type minibot status alive distance 5.2 x 7.43782 y 12.4632 id 4 charge 73 held
false]
```

In plain language, this means: Information at time 64 about the minibot with the name minibot4: It is alive, stays at a distance of 5.2 units from the Minder, has the ID 4 and the load 73 and is not held by another agent.

The conviction management subsystem obtains its information on the one hand from the information of the perception subsystem, and on the other hand from a conviction database, in which it is recorded, for example, that fences are objects with which a ditch can be secured. In order to be able to delete false beliefs from the system, each conviction is assigned a *defeater* . If the *defeater* is evaluated as true, then the associated belief is deleted from the corresponding database. An example:

```
[belief time 20 name minibot8 type minibot status alive distance 17.2196
x 82.2426 y 61.2426 id 8 charge 88 held false
[defeater
[[belief == name minibot8 == x ? Xb y ? Yb ==]
[WHERE distance(myself.location, Xb,Yb) < sensor_range]
[Not New_Sense_Datum == Minibot8 ==]]]]
```

The *defeater* in this case means: "IF I have a belief about minibot8 AND I don't have any new perceptual data about minibot8 AND I'm in a position where I believe I should have new perceptual data about minibot8 THEN my belief is wrong."

The subsystem of reactive plan execution is necessary so that the Minder can react quickly to changing external circumstances. For example, if he has a plan to move from one crèche position to another, then this plan should be carried out without taking up too many resources. To do this, MINDER1 uses a method developed by Nilsson (1994) called *teleo-reactive* (TR) *program formalism* . MINDER1 has thirteen such TR programs, which allow the Minder to search for objects, maneuver in space or head for certain targets, for example.

In order to be able to use TR programs, the Minder first needs targets. These are produced by the subsystem of pre-attentive motive generation, which consists of a series of generactivators. One example is the Gene Activator *G_low_charge*, which scours the belief database for information about babies with a low load. If he finds such information, he forms a motif from it and deposits it in the motif database. An example:

[MOTIVE motive [recharge minibot4] insistence 0.322 status sub]

The status *sub* means that the subject has not yet overcome the filter.

MINDER1 has eight gene activators that express different *concerns*.

9.4.2. Das deliberative Subsystem

The deliberative subsystem of MINDER1 consists of the modules Filter, Motif Management and Plan Execution. All these modules are shallow, so they have little depth of detail.

The filter threshold in MINDER1 is a real number between 0 and 1. A motivator with the status *sub* can overcome it if its insistence value is higher than the value of the filter threshold. The status of the motivator then changes from *sub* to *surfacing*. A motivator that does not succeed in overcoming the filter during a time cycle can be sent into the race again by the Gener Activator with a newly calculated insistence value.

All motivators that have overcome the filter are processed by the Motif Management and receive the status surfaced. Motive management works with the three modules *deciding*, *scheduling* and expanding.

The "Disposition" module determines whether the motivator should be processed immediately or later. If it is processed directly, it is given the status *active*; if it is processed later, it is given the status *suspended*.

Such a decision is usually only possible after the motivator has been subjected to a closer examination. This is done through the "Expansion" module. The motivator is expanded so that it contains a metaplan. An example:

[MOTIVE motive [save ditch1 minibot5] insistence 0.646361 status active]
will be partially expanded to:
[MOTIVE motive [save ditch1 minibot5] insistence 0.646361 status active
plan [[decide] [get_plan]]
trp [stop]
importance undef]

The plan steps *decide* and *get_plan* of the metaplan are calls to management processes.

The motivator can be executed directly from the "Expansion" module (status: *executed*), a saved plan can be retrieved (status: *active, meta*) or the motivator can be passed on to the "Decision" module (status: *active, meta*).

The active motivator is subjected to a routine in the "Decision" module, which determines its importance. This importance is determined heuristically from the insistence value; determining the urgency of a motivator has not yet been implemented in MINDER1. As a result, it can either be rejected (status: *suspended, meta*) or further processed (status: *active, meta*).

Finally, the subsystem of plan execution is implemented in a number of TR programs, which in turn can call other TR programs. MINDER1 has nine *Action Primitives*.

9.4.3. The Meta-Management Subsystem

MINDER1 has two modules in the meta-management subsystem: changing the filter threshold and detecting perturban states.

The filter threshold module observes the number of motivators who manage to overcome the filter. If there are more than three, the threshold is raised until there are only three or fewer motivators in progress. Motivators that have already been processed with the status suspended are also returned to the *status sub* if their insistence value is lower than the threshold value of the filter. Only when there are fewer than three motivators in the process is the threshold lowered again.

The interaction of this module and the gene activators, which recalculate the insistence values of the motivators, leads to a constant movement of the motivators from the pre-attentive to the attentive state and back.

The system is not yet able to generate perturbances in this form. Sloman therefore speaks of "proto-perturbans" in connection with MINDER1. In order to create proto-perturbances, Sloman resorts to a trick that is necessary due to the lack of complexity of the system. Motivators related to damaged minibots (babies) are by definition given a high insistence score; the management processes, however, assign a low degree of importance to these motivators - i.e. a deviation from the usual behavior of the management processes, in which the importance is evaluated analogously to the insistence.

The corresponding module in the meta-management subsystem was designed to measure the *rate of rejection* of motivators. If this value exceeds a certain threshold, then a proto-perturbant state has occurred.

In fact, corresponding proto-perturbances occur in MINDER1. However, the subsystem cannot do anything with this information; the entire system is still too immature for that.

9.5. Summary and evaluation

Sloman's theoretical approach is certainly one of the most interesting in the development of emotional computers. It is not so much his specific interest in emotions, but rather his emphasis on architecture that opens up new perspectives.

Sloman theoretically carries out most consistently what others have also considered: that there is no fundamental difference between emotion and cognition. Both are aspects of control structures of an autonomous system.

However, a detailed examination of Sloman's work from 1981 to 1998 brings to light a number of ambiguities. For example, the distinction between the terms goal, motive and motivator is unclear because he uses them arbitrarily and in constant alternation.

It is also not clear exactly what function perturbances have in the emergence of emotions and how they are related to the global alarm system he postulates. It is interesting that in the early works there was no mention of this alarm system, but mainly of perturbans; in the current works, this relationship has been reversed.

The proof that Sloman wanted to provide with MINDER1 is not convincing in its current form. Neither do perturbances arise from the interaction of the elements of the system (the programmers had to help mightily to create proto-perturbances), nor can further conclusions be drawn about the emotions of people from them.

Nevertheless, it is precisely the theoretical depth and breadth of Sloman's work that can give new impetus to the study of emotions. His combination of design-oriented approach, evolutionary theory, and discussion of virtual and physical machines is deeper than any other approach to the construction of autonomous agents.

10. The libidinal economy of the computer

Ian Wright, a member of Sloman's research group in Birmingham, has further developed his theory into a *computational libidinal economy* (Wright, 1997).

Wright categorizes the theories of Simon, Sloman, Frijda as well as Oatley and Johnson-Laird under the term "*design-based interrupt theories*" and formulates three points of criticism that apply to all of the above-mentioned approaches.

10.1. Criticism of interrupt theories of emotion

10.1. The *control precedence problem*

In his approach, Simon distinguishes between emotions with an interruptive function, which have a high adaptive value, and emotions with a disruptive effect, which tend to run counter to adaptive behavior. According to Wright, the theories criticized have not yet solved the problem of why a disruptive, i.e. non-adaptively meaningful emotion can take control of an intelligent system and maintain it for a longer period of time. Apparently, in such cases, the meta-management system is not able to end the disruption quickly. In order to explain such phenomena, the theories would have to be extended to include phylogenetic and ontogenetic and social aspects.

10.1.2. The *emotional learning problem*

Wright criticizes the present theories for not presenting mechanisms that explain the connection between emotional states and learning processes. For him, emotional states not only have a motivational component, but are also important impulses for learning processes. Frijda (1986) also expressly points this out. Related to this is the correlation between the intensity of an emotion and the learning process, which is not explained by the interrupter theories.

10.1.3. The *hedonic tone problem*

According to Wright, the theories at hand do not explain the mechanisms on which *hedonic tone* signals are based, why such signals are "simple", why they are different from semantic signals, and why, in the case of pleasure and pain, they are either positive or negative.

Simon, Wright argues, simply sweeps feelings under the physiological rug by postulating that all hedonistic states are consequences of the perception of bodily states. Therefore, it is not possible with his theory to explain, for example, a state such as "grief" and the associated psychological pain, which does not necessarily have to be associated with physical states of arousal.
For Frijda, Oatley & Johnson-Laird, and Sloman, hedonistic components are simple, phylogenetically older control signals. This gives them at least a function at the level of information processing.

Frijda emphasizes the importance of the hedonistic coloring of emotional states. His theory postulates relevance signals for joy, pain, wonder, or desire that occur when an event is compared to the satisfaction conditions of different concerns.

Oatley and Johnson-Laird explain the hedonistic components of fundamental emotional states through their concept of control signals. Their theory assumes, for example, that the hedonistic coloring of joy or sadness is caused by fundamental, irreducible control signals. Because of their functional role, control signals have different hedonistic values. The control signal for *sadness* , for example, has the function of canceling or changing plans, while the function of *happiness* is to maintain or follow up on plans.

In Sloman's theory, *insistence* is not associated with hedonistic components. However, Sloman sees the importance of hedonistic components, which play a motivational role as negative or positive evaluations by breaking off or maintaining actions. He admits that his model needs to be extended to include a *pleasure and pain* mechanism.

10.2. The concept of "valence"

Wright tries to find a solution to the latter problem by first proceeding with definition. *Hedonic tone* is too general a term for him. That is why he uses the term "valency".

First of all, Wright differentiates between *physiological* and *cognitive* forms of pleasure and pain. He then states that hedonistic coloring is always associated with a quantitative dimension, the intensity. He cites Sonnemans & Frijda (1994), which distinguish six aspects of emotional intensity: the duration of an emotion, perceived physical changes and the strength of perceived passivity (*loss of control of attention*), memory and re-experience of the emotion, strength and *drasticness* of the tendency to act, and *drasticness* of actual behavior, changes in beliefs (*beliefs* of the) and their influence on long-term behaviour and overall perceived intensity. Wright points out that none of these categories describe the intensity of hedonistic coloring, but that the category of "strength of perceived passivity" is related to it, because both intense pleasure and intense pain are difficult to control at will.

Wright then defines valence as follows:

> “Valency is a form of cognitive pleasure or unpleasure not linked to information concerning bodily locations, and is a quantitatively varying, non-intentional component of occurrent convergent or divergent emotions. Valenced states are contingent on the success or failure of subjectively important goals.”
> (Wright, 1997, p. 115)

Wright explicitly points out that valence, according to his definition, should not be confused with short-term control states of *pleasure* and *unpleasure*, by which ongoing activities are protected or terminated; nor is valence identical with values, i.e. qualitative affective dispositions towards certain states. Valence is *achievement pleasure* or *failure unpleasure*, which occurs when certain concerns that are very important for a system are fulfilled or violated.

10.3. Learning in Adaptive Agent Systems

Wright takes Sloman's system as a basis and extends it to include the reinforcement *learning* (RL) component. In order to be able to implement this mechanism, he first postulates: "A society of mind needs an economy of mind."

For Wright, the aspect that RL always contains a selection component is essential: reinforced actions have a stronger tendency to be repeated than non-reinforced ones.

In order to use RL at all levels of a multi-agent system, a corresponding reward mechanism is required. Wright relies primarily on four corresponding algorithms: Q-Learning, Classification Systems, XCS and Dyna.

10.3.1. Q-Learning

In Q-learning (Watkins & Dayan, 1992), an agent tries to learn, for each possible situation-action combination, what the value of that action is when performing it in the given situation. At the beginning, the values for all possible situation-action combinations are set to a default value. The goal of this system is to update the values so that they lead to the maximum *cumulative discounted reward*.

The maximum cumulative reward at any given time consists of the reward for the immediately following action and the expected rewards for the subsequent actions. These rewards are discounted in such a way that immediately expected rewards are valued higher than expected rewards in the further future.

The reward predictions P for each possible situation-action combination are stored in a two-dimensional matrix. From this table of values, the algorithm selects the action that has the highest predictive value for the current situation. With the help of an update rule, the values are then recalculated.

One of the greatest weaknesses of Q-learning is, among other things, that in the case of large areas of situations and actions, the corresponding tables become excessively large and make an economic search for trial and error impossible.

10.3.2. The classification system of Holland

Holland (1995) has developed an algorithm called *classifier system* . In this way, he wants to ensure that a learning success that insists on a sequence of actions of several modules is also given to all modules involved in the form of a reward.

In his system, there are numerous classifiers that are nothing more than condition-action rules. Some of them observe the environment and, if their own rule is met, send corresponding messages to a kind of message *list*. Other classifiers suggest their specific proposals for action based on the information on the bulletin board. The probability that the system will accept such a proposed course of action is mainly based on the strength of the classifier, which in turn derives from how successful its proposals have been in the past.

If a classifier's accepted proposed action leads to success, he receives a reward that increases his strength. If his suggestion is followed by failure, he receives a punishment in which his strength is diminished. In doing so, he shares the reward or punishment with all the other classifiers who helped him with his proposal.

This *credit assignment* is done via the *bucket brigade* algorithm. The algorithm is called *the bucket brigade* because not only is the last classifier in a series of classifiers rewarded or punished, but the rewards or punishments are distributed proportionally to the classifiers working with it - just as firefighters used to pass the buckets of water along a chain when extinguishing fires. Thus, a reward can be propagated backwards through the system and trigger corresponding reinforcements in certain chains of action.

Holland has also coupled his model with a genetic algortihmus. Successful classifiers are paired and can create new classifiers that can then work even more effectively.

10.3.3. XCS

With XCS, Wilson (1995) introduced a further development of Holland's *classifier system* . XCS addresses one of the weaknesses of Holland's system, in which only the strongest are rewarded. The success of an XCS agent is not determined by its absolute strength, but by its ability to make correct predictions about the probability of success of its actions. So if a classifier in the XCS system correctly predicts that it will receive a low reward, it qualifies it for inclusion in the genetic algorithm.

10.3.4. Dyna

The Dyna architecture by Sutton (1991) goes one step further, because it has the ability to plan. Before an action is initiated, Dyna can play through the consequences of possible actions "in his head" through trial and error within a world model and thus develop an optimized action strategy.

10.3.5. The concept of "value"

Wright points out that RL algorithms are trial and error learners who, in order to be adaptive, receive a reward that is quantitatively staggered. "Unfortunately, the form or forms of value in natural reinforcements learners are unknown." (Wright, 1997, p. 139)

Wright points out that *value* can have two different meanings: First, it is used when an object is valued: Someone values an object very much, it is dear to him. The other use is to attribute value to an object with a view to a specific goal: a chainsaw usually has a higher value than an axe for a lumberjack.

Wright distinguishes between the value that an external object can have and the value that an internal state of a system can have. For Wright, value is a relationship between a purposeful system and its own internal components. *Value* "*refers...to the utility of internal substates*" (Wright, 1997, p. 138).

Value is both a scalar quantity and a control signal. The form that *value* takes in RL algorithms is that of a scalar quantity. Such a scalar quantity, in contrast to a vector, cannot be decomposed into components with different semantics. *Values* specify a besser_als relationship between *substates* and have no meaning beyond that.

In an RL system, the values of the different *substates* change over time; *value* thus controls the alternative course of action to be carried out in each case. The value of a *substate* is that *it can be used to buy* processing power.

10.4. Wrights *currency flow hypothesis*

Wright points to the coordination problem in multi-agent systems (MAS), to which Oatley (1992) has already drawn attention. This is especially true for Adaptive Multi-Agent Systems (AMAS). For Wright, the solution to this is an internal economy with a *currency flow*.

Wright compares an AMAS to a business society:

> "In the abstract, economic systems are selective systems: the trials are the various concrete labours that produce commodities, the evaluatory mechanisms are the various needs and demands of individual consumers, and selection occurs through the buying and selling of commodities. Over time what is produced matches what is required given available resources."
> (Wright, 1997, p. 154)

Based on this, Wright develops his *currency flow hypothesis* (CFH):

> "**The currency flow hypothesis (CFH) for adaptive multi-agent systems:** Currency flow, or *circulation of value*, is a common feature of adaptive multi-agent systems. Value serves as a basis for coordination; it integrates computational resources and processing by constraining the formation of local commitments. Circulation of value involves (i) altering the dispositional ability of agents to gain access to limited processing resources, via (ii) exchanges of an explicitly represented, domain-independent, scalar quantity form of value that mirrors the flow of agent products. The possession of value by an agent is an ability to buy processing power."
> (Wright, 1997, p. 160)

10.5. The CLE system in detail

Wright's *computational libidinal economy* combines Sloman's model of an intelligent system with a learning mechanism and a motivational subsystem that maintains emotional relationships with other agents. In doing so, Wright also hopes to solve a problem of Sloman's model, which he calls the *valenced perturbant states problem*, because it cannot explain how perturbances with a valencied component come about.

Wright begins the description of his model by specifying CFH again for natural RL:

> "**The currency flow hypothesis for natural reinforcement learners (CFHN):** The currency flow hypothesis holds for the reinforcement learning mechanisms of individual, natural agents that meet a requirement for trial and error learning."
> (Wright, 1997, p.163)

The description of CLE includes several aspects: a libidinal selective system, a scalar quantity form of *value*, credit allocation, and a value circulation theory of *achievement pleasure* and *failure unpleasure*.

10.5.1. The libidinal selective system

Wright's libidinal selective system is a cognitive subsystem whose main task is to develop social relationships. It contains the following components:

1. Untaught *conditions of satisfaction:*
   These are innate satisfaction mechanisms that have been selected by evolution and specify fundamental *attachment goals*, such as orgasm, positive emotional signals from the opposite sex, etc. According to Wright, evolution is therefore also the cause of *attachment motivation*.
2. Means *of satisfaction:*
   These are *motivational substates* or agents that constitute the means of satisfying the different *attachements goals*. They, in turn, can produce motivators for higher levels.
3. Learnt *conditions of satisfaction:*
   These are learned satisfaction mechanisms that have inherited their reinforcement mechanisms from innate satisfaction mechanisms and may be able to dominate them.
4. A *selective cycle: As a*
   selective system, the libidinal system fulfills three functions: it generates *substates*, which represent possible mechanisms of satisfaction; it evaluates these *substates*; it selects and deselects *substates*. This is done by the reinforcement mechanisms described above.
5. Substate *discovery:*
   The libidinal system produces new substates through its genetic algorithm, which consist of new agents, new rules, etc., and evaluates and selects them accordingly.
6. Varieties *of control substates:*
   The control structure within the libidinal system is not static, but dynamic. Due to the constant selective processes, certain *substates* can move up in the hierarchy and others down. The net effect is one of diffusion, where a strong state of control spreads through the entire system into numerous *substates* and can sometimes even become an automatic reaction. Among these *substates,* Wright also includes the *libidinal generactivators,* which produce motivators for attentive *processing* and which for him correspond to Frijda' *s concerns*.

10.5.2. The *conative universal equivalent* (CUE)

In Wright's model, CUE represents the scalar quantity form of value that he demands . He uses the term "conative" here in the sense of "motivational". CUE is the universal medium of exchange between the *substates* of the libidinous system. Possession of CUE means the ability *to buy processing power*. This can take various forms:

a. The dispositional ability to engage pre-attentive processing resources;
b. the dispositional ability to produce motivators for management processing;
c. the dispositional ability to become aware of motivators and to command management resources.

Thus, CUE is causally related to the interruptive abilities of motivators and their ability to claim attentional resources.

10.5.3. *Credit assignment*

The exchange of CUE reflects the flow of semantic products in the system: in order to enter the cycle, a *substate must pay* the *substate* that delivered the semantic product to which the first *substate* responds. This distribution of CUE to preceding *substates* is done according to Holland's bucket brigade algorithm.

Other aspects of the credit allocation system are:

a. Amplifier as a source of CUE (*derivation of CUE from reinforcers):*
CUE is only allocated if it meets the satisfaction conditions of the innate or derived learned reinforcers.
b. Gain of CUE*:*
Substates can increase their CUE value (positive gain).
c. Loss of CUE (*loss of CUE):*
Substates can lose CUE (negative gain).
d. Accumulation *as reinforcement:*
The accumulation of CUE by a *substate* represents RL.
e. Loss *as deselection:*
The loss of CUE by a *substate* represents its partial deselection.
f. CUE as *internal economy with control semantics:*
CUE is a *domain-independent* control signal that does not refer to other things inside or outside the system.

10.5.4. The Value Circulation Theory

CLE has two distinguishable internal states: intentional and non-intentional. The intentional component of the CLE is the set of substate products, especially the motivators produced by the libidinous *generactivators*. These have a representational content, they revolve "around" something. The non-intentional component of CLE is the *circulation of value*. This value circulation is a flow of control signals, not semantic signals.

The value circulation requires a module of the overall system that observes and registers the internal flow of CUE; i.e. the meta-management layer called by Sloman. This mechanism will detect movement of CUE in the system at any point in time. For each *substate* , the values change depending on whether it is rewarded (positive) or punished (negative).

Wright uses a thought experiment to illustrate what this can lead to. A virtual frog (*simfrog*) learns to catch flies in a virtual environment. If the necessary *substates* are successful, the meta-management layer records an increase in CUE compared to a previous point in time. Suppose that the observations of the meta-management layer are linked to the skin color of the frog: positive values lead to yellowing of the skin, negative values to blueing and no changes to any skin change. After a successful fly catch, the frog would notice a change in its skin color that it cannot explain. At the same time, has either positive or negative sensations of varying intensity (depending on the change in the CUE state). A non-intentional state of control arises, triggered by the circulation of value in a system with a meta-management layer.

Wright therefore adds another element to his libidinal economy: valency as the *monitoring of a process of credit assignment.* The registration of value circulation produces valencied states that represent a form of cognitive *achievement pleasure* or *failure unpleasure* .

a. Negative valence means a loss of CUE: A registered circulatory process that involves a loss of value corresponds to negative valence.
b. Positive valence means an increase in CUE: a registered circulatory process that includes an increase in value corresponds to positive valence.
c. Intensity is the measure of the exchange of CUE: The exchange rate of CUE between *substates* corresponds to the quantitative intensity of the valenced state.
d. Gain of CUE is related to the achievement of goals: If the achievement of a goal coincides with the satisfaction conditions of an amplifier, there can be a gain in CUE.
e. Loss of CUE is related to the failure to achieve goals: If the failure to achieve a goal is consistent with the satisfaction conditions of a negative reinforcer, there may be a loss of CUE.

> “In other words, certain types of `feelings' are the self-monitoring of adaptations; that is, the pleasure and unpleasure component of goal achievement and goal failure states is the monitoring of a movement of internal value that functions to alter the dispositional ability of substates to buy processing power and determine behaviour.”
> (Wright, 1997, p. 176)

10.6. CLE in Practice

Wright illustrates the functioning of his model using the example of grief. From an analysis of statements by mourners, he picks out a number of phenomena and tries to explain the underlying processes with the help of his theory.

1) The repeated and continuous interruption of attention by thoughts about and memories of the deceased.
If a bonding structure to X exists, then motives and thoughts related to it will emerge and successfully compete for processing resources of attention. These cyclical processes can include the wish that the deceased may still be alive or the wish that something could have been done to prevent his death. Therefore, due to the news of X's death, these and other *substates* will very likely circulate through the system in which they are deeply rooted due to the intense bond with the deceased. The agent's thought processes are shaken by perturbances and partially escape his conscious control.

> “The structure of attachment explains why motives relating to X are likely to disrupt attention. (a) X-related motives will be given high insistence values because the relationship with X is strongly positively rewarded, and therefore important, and X has suffered great harm. (b) Meta-management control processes ensure that motives and thoughts pertaining to X are always decided as soon as possible, so that such motives tend to grab attentive resources immediately. (c) Dedicated evaluation procedures rate X-related motives preferentially, assigning skewed importance, urgency and cost-benefit measures. (d) Predictive models, triggered by X-related motives, will consume computational resources by attempting to reason about X's needs and possible reactions to things. (e) In a resource-limited system, the proliferation of motives pertaining to X may `crowd out' other motive generators.”
> (Wright, 1997, p. 201)

2) The difficulty of accepting the death of the deceased.

Updating a large database and propagating the information through the system takes time. In addition, the agent has affective reasons for not accepting the information, because it would mean that years of development work may have been in vain. Finally, there is the agent's knowledge of the long and painful grieving process, which he would like to postpone.

3) The disruptive effect on everyday functioning.
Daily goal processing is complicated by *management overload*, which can be traced back to the disruption of the motif management processes.

4) Periods of relative normality in which grief is pushed into the background.
For important new tasks, the filter threshold is set so high that thoughts of the deceased cannot get through. After completing the task, the filter threshold drops again, and mourning occurs again.

5) Try to fight grief.
The activity of a metaprocess that notices the disorder and tries to fight it. However, this rarely succeeds; often the result is only a pushing of the motivators below the filter threshold, where they increase in urgency and wait for the filter threshold to drop. Then they increasingly penetrate and lead to a loss of control of the agent over the system. The perturbances will only decrease when the CLE *has largely completed* the process of detachment.

6) Second-order motivators, e.g. evaluation of grief.
Metamanagement processes that are strongly influenced by culture.

7) The subjectively experienced pain.
Loss of CUE leads to negative states that are experienced as pain. Negative valence prevails because *generactivators* produce motivators that can no longer be satisfied. This leads to overproduction, which leads to gradual deselection and has a high negative valence.

8) Crying.
If motives repeatedly penetrate through the filter that disrupt the normal management process and it is not possible to suppress them for a longer period of time, an agent often no longer comes up with a strategy to change the situation. "*Crying is the plan of last resort*, and can be triggered by negatively valenced perturbant states." (Wright, 1997, p. 207)

10.7. CLE and Problems of Interrupt Theories

Wright claims to have found a solution to the problems of interrupter theories of emotion that he has outlined with his model. He explains this for the four problem areas mentioned.

10.7.1. CLE and the *Hedonic Tone Problem*

In their theory, Oatley and Johnson-Laird postulate fundamental and irreducible control signals for emotions such as *happiness* and *sadness*. In CLE, one element is sufficient for this: value circulation consists of simple control signals that are observed and registered by another entity. Depending on the outcome of this process, the emotions arise, for which Oatley and Johnson-Laird assume two separate signals.

Value circulation also has the advantage of coordinating a variety of relatively autonomous *substates* . The task of value circulation is ultimately only to attribute positive or negative balances. All other effects are second-order and result from the original, simple function.

According to Wright, the CLE theory also explains why control signals differ from semantic signals. *Value* is nothing more than a means of establishing besser_als relationships between *substates* and thus does not contain any semantic content such as beliefs or desires.

10.7.2. CLE and the *emotional learning problem*

By introducing a fictitious currency CUE and circulating it through the system, learning effects become possible. Reinforcement learning can thus change the abilities of gene activators to interrupt processing processes and claim the system's resources for themselves.

Emotions also have a strong influence on learning processes. The more results of behavior are accompanied by positive or negative feelings, the better the corresponding behavior is learned or avoided. By gaining CUE, *substates* gain power in the system; the more CUE they have, the stronger the registered intensity of the valenced state.

10.7.3. CLE and the *valenced perturbant states problem*

Wright equates a concern in Frijda's sense with a libidinal gene activator, whose strength is defined by how much processing capacity it can buy. At the same time, this also determines his disposition to be able to influence behavior. This strength is based on the CUE he accumulates.
Gene activators of the libidinal system, which have a lot of CUE, produce motifs with a high potential for interruption. A high gain or a large loss of CUE leads to a valenced state, which can also be accompanied by a loss of control (grief, triumph).

> “.. occurrent reinforcement learning together with the monitoring of credit assignment plus loss of control of attention is experienced as a valenced perturbant state.”
> (Wright, 1997, p. 183)

10.7.4. CLE and the *control precedence problem*

Why can dysfunctional and non-adaptive emotions take control and not be pushed back through the meta-management layer? Wright offers as an explanation that the process of accumulation of CUE by libidinous gene activators cannot be controlled by this layer; it can only register the processes. Only the libidinal selective system itself can take away the strength of a *substate* that has an excess of CUE and thus has a disruptive effect on the overall system. Only when this has been done will the state of loss of control be lifted.

10.8. Summary and Evaluation

Wright's model attempts to solve a number of problems that have been circumvented in other computer models. Of particular interest is his proposal for the treatment of the *hedonic tone problem*. While other models always define the hedonistic value of an event directly, Wright tries to model it as a property of a system.

The combination of Sloman's theoretical approach with reinforcement learning and the introduction of an imaginary currency whose circulation through the system is responsible for emotional processes requires a model of high complexity, but offers a conclusive explanatory approach for the emergence of emotions as well as for disruptive emotional processes within the framework of the model - and not only on an abstract level, but already close to operationalization.

On the other hand, with Pfeifer against Wright, one could bring up the accusation of "overdesign". The already complex Sloman model, which was implemented in MINDER1, becomes several degrees more complex due to Wright's additions and thus places high demands on the programming of the system and the underlying computer capacity.

Of all the models presented, Wright's is the only one that neither excuses "emergence of emotions" as a reason for a lack of integration into a model, nor does it firmly program emotions from the outset. It remains to be seen to what extent his attempt at a theoretical justification of emotions in connection with a "partially emergent" design will prove to be valid or not when implemented in an actual model.

11. A new paradigm?

The approach of viewing emotions as a feature of the architecture of an intelligent system has led to increased interest in this topic in recent months. While in 1997 only a total of two papers on the topic of "Agents" were presented at the leading world congresses on the topic of "Agents", in 1998 there was already a first congress exclusively on the topic of "Intelligent Agents"

A number of researchers have actually begun to develop emotional autonomous agents based on the principles prescribed by Simon, Toda and also Sloman. Many of these approaches are still at the stage of theoretical exploration; Some have already been rudimentarily implemented.

A fundamental commonality of all agent-centered approaches is the conception of emotions as control signals in an architecture, which must possess a system that moves independently in an uncertain environment. The function of emotions is to focus the system's attention on an external or internal aspect that has a meaning for essential goals or concerns of the system, thus assuring it processing priority.

11.1. Velásquez's Models

11.1.1. Cathexis

Velásquez (1997) has developed a model based on the "Society of Mind" theory of Minsky (1985). He calls it *Cathexis*, a term he defines as a "concentration of emotional energy on an object or idea" (Velásquez, 1997, p.10).

In his model, emotions consist of a variety of subsystems:

> “Emotions, moods, and temperaments are modeled in Cathexis as a network of special emotional systems comparable to Minsky’s “proto-specialist” agents (...) Each of these proto-specialists represents a specific emotion family... such as *Fear* or *Disgust*.”
> (Velásquez, 1997, S. 10)

Each of these proto-specialists has four types of sensors that are responsible for measuring internal and external states: neural sensors, sensorimotor sensors, motivational sensors, and cognitive sensors. In addition, each Proto-Specialist is characterized by two thresholds, which Velásquez refers to as Alpha and Omega: Alpha is the threshold above which the activation of the respective Proto-Specialist begins; Omega is the saturation limit of a proto-specialist. Finally, each proto-specialist has a *decay function* that affects the duration of its activation.

In his model, Velásquez distinguishes between *basic emotions* and *emotion blends/mixed emotions*. In defining *basic emotions*, he relies on Ekman and Izard and defines them as follows:

> “In this model the term basic... is used to claim that there are a number of separate discrete emotions which differ from one another in important ways, and which have evolved to prepare us to deal with fundamental life tasks...”
> (Velásquez, 1997, S.11)

Die *basic emotions* in Cathexis sind *Anger*, *Fear*, *Distress/Sadness*, *Enjoyment/Happiness*, *Disgust* und *Surprise*.

Emotion blends or *mixed emotions* are emotional states that arise when several different emotional proto-specialists representing the *basic emotions* are active without one of them dominating the others.

Finally, the model also knows *moods*, which differ from emotions only by the level of arousal.

Emotions in *Cathexis* are evoked by cognitive and non-cognitive *elicitors*, which come from the same categories as the system's sensors. The *cognitive elicitors* for the *basic emotions* are based on a modified version of Roseman's emotion model.

In Velásquez's model, the intensity of an emotion is influenced by several factors:

> “Thus, in Cathexis, the intensity of an emotion is affected by several factors, including the previous level of arousal for that emotion..., the contributions of each of the emotion elicitors for that particular emotion, and the interaction with other emotions...”
> (Velásquez, 1997, S. 12)

The behavioral repertoire of the system has three essential elements: an *expressive component*, with the help of which it communicates its current emotional state, consisting of face, body and voice; an *experiential component*, which learns from experience and affects the motivations and action tendencies of the system; and an action selection mechanism, which is based on the calculated *behavior values* of different alternative courses of action, the one with the highest value.

The system regularly goes through so-called *update cycles*, in which the following cycle is handled:

> “1. Both the internal variables (i.e. motivations) and the environment are sensed.
> 2. The values for all of the agent’s motivations (both drives and emotions) are updated....
> 3. The values of all behaviors are updated based on the current sensory stimuli (external stimuli and internal motivations).
> 4. The behavior with the highest value becomes the active behavior. Its expressive component is used to modify the agent’s expression, and its experiential component is evaluated in order to update all appropiate motivations.”
> (Velásquez, 1997, S. 13f.)

Velásquez has implemented *Cathexis* in the form of a computer model that he calls "Simón the Toddler". The screen shows the face of an infant capable of different emotional expressions and rudimentary verbalizations. The user interacts with the system by, for example, changing the parameters of Simón's proto-specialists, varying the level of neurotransmitters or interacting directly with it by feeding, petting it, etc.

Currently, the model has 5 drive proto-specialists (hunger, thirst, temperature regulation, fatigue, interest) and a repertoire of 15 behavioral alternatives, including sleeping, eating, drinking, laughing, crying, kissing, and playing with toys. These are to be expanded step by step in the course of further model development.

11.1.2. Yuppy

Yuppy is a robot that represents an emotional pet. It is a further development of the *Simón the Toddler* model. Yuppy was first developed as a virtual simulation before being given a body. Velásquez calls it an example of a system with *emotion-based control*.

The model is composed of a set of *computational units* consisting of three main components: an input, an estimation mechanism, and outputs. An essential part of the assessment mechanism are the *releasers*. They filter sensory data and identify particular conditions on the basis of which they then send excitatory or inhibitory signals to the subsystems connected to them.

Velásquez follows Damasio and LeDoux and distinguishes between natural and learned *releasers*. *Natural Releasers* are *hard-wired*; *Learned releasers* are learned and represent stimuli that are associated with the occurrence of *natural releasers* or can predict their occurrence. In the language of other models, *natural releasers* correspond to primary emotions, while *learned releasers* are identical to secondary emotions. The latter require more processing capacity and are more complex, as they are based, among other things, on personal emotional memories that need to be activated.

In Yuppy's work, drives are motivational systems that drive the agent to action. Drive systems are clearly distinguished from emotional systems.

Yuppy's emotional systems represent six groups of basic affective reactions: *anger*, *fear*, *distress/sadness*, *enjoyment/happiness*, *disgust,* and *surprise*. Velásquez distinguishes between cognitive and non-cognitive *releasers* of emotions. He distinguishes between four groups:

a. The neural group includes the effects of neurotransmitters, brain temperature, and other neuroactive agents that can lead to an emotion and that are influenced by hormones, sleep, diet, and environmental conditions.
b. The sensorimotor group includes sensorimotor processes such as facial expressions, posture and muscle potential, which can not only regulate existing emotions but also evoke emotions.
c. The motivational group includes all the motivations that lead to an emotion.
d. The cognitive group includes all kinds of cognitions that activate emotions, e.g. assessments of events, comparisons, attributions, wishes, beliefs or memories.

Yuppy's perception system consists of two color CCD cameras as eyes; a stereo audio system with 2 microphones as ears; infrared sensors for obstacle detection; an air pressure sensor to simulate touch; a pyrosensor that detects changes in room temperature when people enter the room, as well as a simple proprioceptive system.

Yuppy's drive system includes four drives: charge regulation, temperature regulation, fatigue and curiosity. Each of these pinions controls an internal variable associated with it, which represents the battery's state of charge, the level of temperature, the amount of energy, and the agent's level of interest.

Yuppy's emotion generation system consists of emotional systems with *natural releasers* for the basic emotions. Velásquez divides emotional systems into three groups:

- *Interactions with Drive Systems:* Unsatisfied urges produce *distress* and *anger*; oversatisfied urges produce *distress*, and instinctual gratification produces *happiness*.
- *Interactions with the environment:* All objects with pink color produce *happiness* to varying degrees; yellow objects produce *disgust*. Darkness creates *fear* and loud noises *surprise*.
- *Interactions with People:* People can pat and discipline Yuppy. This creates either pleasure or pain. Joy leads to *happiness*; Pain produces *distress* and *anger*.

Yuppy's behavioral system consists of a distributed network of about 19 different types of behavior, which are primarily concerned with satisfying his needs and interacting with people. Examples of such behavior are "*search for bone*", "*approach bone*", "*recharge battery*" or "*approach human*".

Like the drive systems and the emotional systems, Yuppy's behavioral systems also have their own *releasers*.

The user can control Yuppy's affective style by manipulating parameters such as thresholds, inhibitory or excitatory compounds, etc. It can also offer the robot internal and external stimuli. Velásquez describes the result as follows:

> “Using the model described before, both the simulated and physical Yuppys will exhibit emotional behaviors under different circumstances. For instance, when the robot’s *Curiosity* drive is high, Yuppy wanders around, looking for the pink bone which people may carry. When it encounters one, the activity of the *Happiness* Emotional System increases and specific behaviors, such as *“wag the tail”* and *“approach the bone”* become active. On the other hand, as time passes by without finding any bone, the activity of its *Distress* Emotional System rises and appropriate responses, such as “droop the tail”, get executed. Similarly, while wandering around, it may encounter dark places which will elicit fearful responses in which it backs up and changes direction.”
> (Velásquez, 1998, S. 5)

Yuppy is also able to learn secondary emotions, which are stored as new or modified cognitive *releasers* . For example, if a human has a bone in his hand and gets Yuppy to come over and get it, he can pet or discipline him afterwards. Depending on the experience, Yuppy produces a positive or negative emotional memory with regard to people, which in turn influences his subsequent behavior.

11.2. The Model of Foliot and Michel

Foliot und Michel (1998) definieren Emotionen als ein “evaluation system operating automatically either at the perceptual level or at the cognition level, by measuring efficiency and significance” (Foliot und Michel, 1998, S. 5). Für sie bilden Emotionen die Grundlage jeder Kognition. Ziel ihres Modells ist es daher, zu zeigen, “how emotion based structures could contribute to the emergence of cognition by creating suitable learning conditions” (Foliot und Michel, 1998, S. 1).

The model was implemented in a virtual Khepera robot. Khepera is a miniature robot model that has a number of sensors and can be expanded with additional components depending on requirements. The "Webots Simulator" not only simulates a Khepera; Programs developed with Webots can be transferred directly to a Khepera.

The environment of the virtual Khepera consists of a city with buildings, a river and green areas. Each of these elements has a specific color. The robot has to move through the city and learn to avoid different types of obstacles.

Foliot and Michel represent emotions on two levels. The *level of process* can evaluate stimuli and trigger different emotions; the *level of state* can provide information about the system.

For Foliot and Michel, the basis of their first experiment was the assumption that an emotion is characterized by a reaction to a positive or negative signal. The model consists of four components:

1. A reflex structure that results in a motor movement in the opposite direction to an obstacle detected by an infrared sensor.

2. An association matrix between the motor behavior and the input of the infrared sensors, the initial value of which is set to zero.
3. A signal that produces an association in the matrix every time the robot encounters an obstacle.
4. A behavioral system with the three alternatives: (a) straight-ahead movement when there is no obstacle in the way and no learned association is active; (b) the triggering of a reflex behaviour when hitting an obstacle; (c) Association of a motor configuration with a known sensory pattern.

The experiment showed that the robot gradually collided less and less with obstacles, but could never move completely flawlessly. To check whether improving the learning system through affective signals would produce better results, the authors conducted a second experiment.

The second experiment is based on Scherer's theory of emotion. It consists of five components:

1. A linear evaluation system that corresponds to Scherer's SEC and in which each evaluation stage is used for the subsequent evaluation.
2. Two systems of states, one of which represents the assessment of a situation, the other the physical body.
3. Two cognitive processes, one of which is responsible for attention selection, the other for deciding on the next movement. The state systems can directly influence these processes.
4. A database of goals.
5. A sensorimotor, a schematic and a conceptual level. The schematic level produces associative schemata between significant patterns and actions.

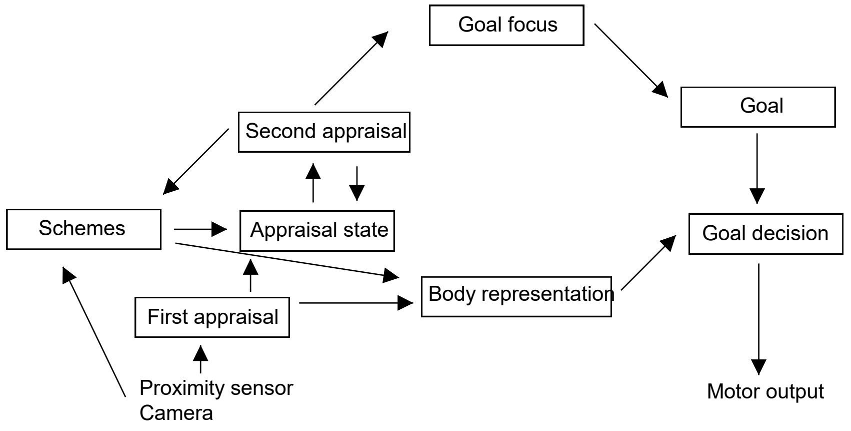


Fig. 17: Controller model of Foliot and Michel (after Foliot and Michel, 1998, p. 4)

The model distinguishes between cognitive and emotional processes. Each emotional process is defined by an assessment sequence that classifies stimuli according to the criteria *novelty*, *pleasantness*, *goal significance* and *coping* . Each stage of this process uses the results of the previous stages as input. *Coping* knows the alternatives "possibility of reacting" and "no possibility of reacting".

The cognitive processes know a primary goal (forward movement) and four secondary goals (left turn, right turn, follow left wall, follow right wall). Each goal is defined by a value in the body representation.

In this model, learning always happens when the average state of the system contains a strong *displeasure* value:

> “This produces a new scheme containing the newer stimulus as a sensory input. The process then waits to observe which goal is associated to this stimulus and [to] check whether this goal allows to come back to a normal state. If this normal state is reached within a small amount of time, the representation is associated to the scheme, otherwise, the scheme is destroyed.”
> (Foliot and Michel, 1998, p. 5)

A central component of the model is the mechanism that produces schemata. The conduct of the experiment showed that when obstacles were avoided, this took place either at the sensorimotor level, when an obstacle was detected by the infrared sensors, or at the schematic level, when an obstacle was not perceived. The schematic level corresponds to a temporary change of destination, which the authors interpret as the result of a danger signal or an internal assessment process.

With regard to the learning process, the system showed two fundamental instabilities in its behavior: Either the robot insisted on its once chosen goal or it changed its goals non-stop. Nevertheless, Foliot and Michel conclude that their approach is fundamentally correct, but requires a more detailed definition of the individual components.

11.3. The model of Gadanho and Hallam

Gadanho and Hallam have investigated the role of emotions in an autonomous robot that adapts to its environment through reinforcement learning (Gadanho and Hallam, 1998). To do this, they worked with a simulated Khepera robot.

The authors modeled their emotion model according to the *somatic marker hypothesis* proposed by Damasio (1994). Damasio assumes that emotions evoke special body feelings. These body sensations are the result of experiences with internal preference systems and external events, and help predict outcomes of certain scenarios. In this way, *somatic markers* are intended to help people make quick decisions without requiring a high processing capacity and a long time.

The model developed on this basis by Gadanho and Hallam knows four basic emotions: *Happiness*, *Sadness*, *Fear* and *Anger*. The intensity of each emotion is determined by the robot's internal *feelings*. These feelings are: *Hunger*, *Pain*, *Restlessness*, *Temperature*, *Eating*, *Smell*, *Warmth* and *Proximity*. Each emotion is defined by a set of constant emotional dependencies and a bias value. For example, the intensity of *sadness* is high when *hunger* and *restlessness* are high and the robot is not eating.

In Gadanho and Hallam's model, each emotion tries to influence the *body state* in such a way that the resulting body state is similar to that which evokes that specific emotion. To do this, the emotion uses a simple hormonal system. A hormone is associated with every feeling. The intensity of a feeling is not derived directly from the value of the body perception that evokes the feeling, but from the sum of the perception and the hormone value:

> “The hormone values can be (positively or negatively) high enough to totally hide the real sensations from the robot's perception of its body. The hormone quantities produced by each emotion are directly related to its intensity and its dependencies on the associated feelings. The stronger the dependency on a certain feeling, the greater quantity of the associated hormone is produced by an emotion.”

(Gadanho and Hallam, 1998, p. 2)

The hormone levels can rise quickly, but slowly subside again, so that the emotional state is maintained for a while, even if the emotion-triggering situation is long over.

The robot equipped with this emotion system has the task of seeking out food sources scattered in its environment and consuming energy. The faster it moves, the more energy it consumes. The food sources consist of lights that the robot can perceive. In order to draw energy from it, he has to push the food source. This releases energy for a short time and also an odor that the robot can perceive. To absorb the energy, the robot must turn and turn its back on the food source. After a short time, the food source is empty and needs a certain amount of time to regenerate. So the robot has to look for other food sources. If a food source has no energy, its light goes out.

In the context of this task, the emotional dependencies on feelings are as follows:

- The robot is happy (*Happy*) if the current situation does not present any problems. He is especially happy when he has used his engines a lot or is just recharging his batteries.
- The robot is sad (*Sad*), when it has little energy and does not absorb any energy.
- If the robot hits a wall, the pain it feels makes it anxious (*fearful*).
- If the robot stays in one position for too long, it becomes restless. This makes him angry (*Angry*). The anger persists until it moves or changes its current actions.

The system learns through *reinforcement learning*. In order to shorten the learning process, the basic behaviors of the robot were programmed in advance so that the system could concentrate on learning behavioral coordination. The three basic behaviors of the robot are avoiding obstacles, seeking out light sources and driving along a wall.

The system has a controller with two separate modules. The *Associative Memory Module* is a neural network that associates the robot's feelings with the currently expected values of each of its three behaviors. The algorithm used is *Q-Learning*. The *Behaviour Selection Module* makes a stochastic selection based on the information from the other module as to which behaviour should be executed next.

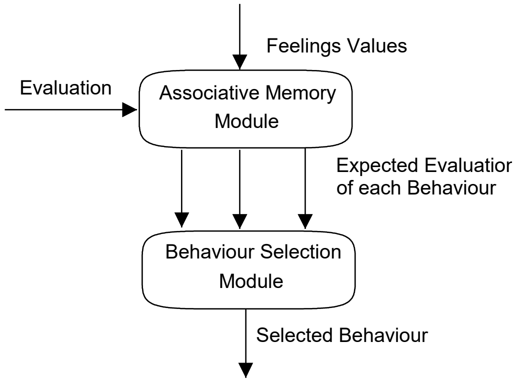

Feelings Values
Evaluation
Associative Memory Module
Expected Evaluatior of each Behaviour
Behaviour Selection Module
Selected Behaviour

Fig. 18: Adaptive controller (according to Gadanho and Hallam, 1998, p. 3)

According to the authors, rewarding or punishing an autonomous robot poses a particular problem. From second to second, the environment or the internal state of the robot changes. If all information is analyzed and behaviors are changed at each transition, this would not only cost immense processing capacity, but would also not provide the robot with any feedback as to whether a chosen behavior might only lead to success after a series of transitions. On the other hand, he must also be able to change dysfunctional behavior quickly. This is where emotions come into play: Your task should be to determine these *state transitions*.

To test this hypothesis, the authors developed a controller with emotion-dependent event detection. An event is detected when one of three conditions occurs:

- there is a change in the dominant emotion;
- the value of the currently dominant emotion differs statistically significantly from the values recorded since the last state transition;
- a limit of 10,000 steps is reached.

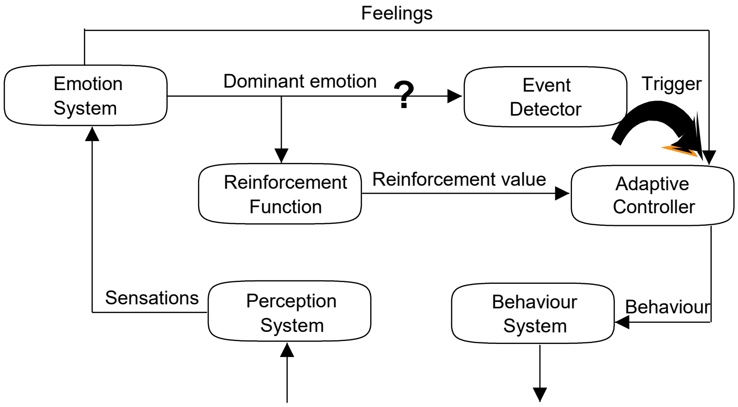


Fig. 19: Emotions and control (according to Gadanho and Hallam, 1998, p. 4)

To test the effectiveness of this event-driven controller, the authors developed three additional controllers:

- Regular intervals (*regular intervals*) - the adaptive controller is triggered every 35 steps.
- Handmade (*hand-crafted*) - all behaviors are hard-coded into the controller, the system can't learn anything.
- Random selection (*random selection*) - the controller selects a new behavior at each step.

Each of these four controllers went through an identical experimental setup. It consisted of thirty different experiments with three million learning steps. In each experiment, a fully charged robot was placed at a randomly selected starting position. For the purpose of evaluation, units of 50,000 steps each were evaluated and data were collected on the following variables:

- the average of the reinforcement received during all steps;
- the average of the gain during the steps in which the adaptive controller was triggered;
- the average of the robot's energy level;
- the number of collisions;
- the frequency of triggering the adaptive controller.

The result is as follows:

Controller	Reinforcement	Event reinforcement	Energy	Collisions (%)	Events (%)
Hand-crafted	0.34	-0.03	0.83	3.0	6.15
Event-driven	0.24	0.04	0.63	0.6	0.52
Regular intervals	0.24	0.20	0.62	1.7	2.86
Random selection	-0.38	-0.38	0.02	5.6	100

Tab. 9: Results of the Gadanho and Hallam experiments (according to Gadanho and Hallam, 1998, S. 5)

According to the authors, the results show that the learning controllers have fulfilled their task. Their energy level is significantly lower on average, but does not reach a critical value. Between the two learning controllers, the main difference is the number of collisions - this is where the event-driven controller is better.

Overall, the event-driven controller is not significantly better than its competitor, but it achieves its learning success with a significantly lower number of events and thus saves much more time.

The authors conclude that the experiments have confirmed their hypothesis about the role of emotions in reinforcement learning.

11.4. The Model of Staller and Petta

Staller and Petta developed the TABASCO architecture, an acronym for "Tractable Appraisal-Based Architecture for Situated Cognizers" (Staller and Petta, 1998). TABASCO is largely based on Scherer's theory of emotion and has not yet been implemented in a simulation.

Staller and Petta conceive of emotions as processes that relate to an agent's interaction with his environment. "In particular, TABASCO models the appraisal process, the generation of action tendencies, and coping." (Staller and Petta, 1998, p. 3)

The basic idea of TABASCO is that the levels of the emotional system postulated by Scherer (sensorimotor, schematic and conceptual) are valid not only in terms of assessments, but also in relation to the generation of action. The two main components of architecture, *perception and appraisal* and *action*, are therefore designed as hierarchies with three levels.

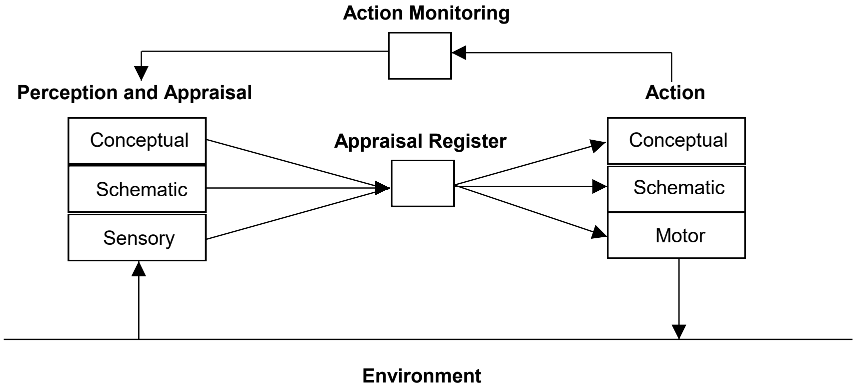


Fig. 20: The TABASCO architecture (according to Staller and Petta, 1998, p. 4)

The *Perception and Appraisal* component: The sensory layer consists of feature detectors for detecting, for example, sudden, intense stimuli or the quality of a stimulus (e.g. *pleasantness*). The schematic layer compares the input with schemata, especially with social and self-schemata. The conceptual layer can think and inference abstractly on the basis of propositional knowledge and beliefs.

The *Action* component: The motor layer contains motor commands. The schematic layer contains tendencies of action and what Frijda calls "flexible programs" (Frijda, 1986, p. 83). The conceptual layer is responsible for *coping*.

Between these two components moderates the *Appraisal Register*, which goes back to a proposal by Smith et al. (1996). It detects and combines the assessment results of the three layers of the *perception and appraisal* component and, based on the assessed state, influences the *action* component.

Finally, the *Action Monitoring* component observes the planning and execution processes of the *Action* component and transmits the results to the *Perception and Appraisal* component, where they are integrated into the assessment process.

Staller and Petta describe their system as a *situated cognizer*. In doing so, they want to underline the importance of both components for an autonomous system. They define *cognizing* (a term first proposed by Chomsky) as "having access to knowledge that is not necessarily accessible to consciousness" (Staller and Petta, 1998, p. 5).

11.5. The Model of Botelho and Coelho

Botelho and Coelho, in the context of their *Salt & Pepper* project, define emotion as "a process that involves appraisal stages, generation of signals used to regulate the agent's behavior, and emotional responses" (Botelho and Coelho, 1997, p.4). The *aim of Salt & Pepper* is to define an architecture that contains mechanisms that play the same role for autonomous agents as the mechanisms that make humans so successful.

The starting point of her considerations is the classification of emotions in a multidimensional matrix "that may be used with any set of emotion classification dimensions" (Botelho and Coelho, 1997, p. 4).

Dimension of classification	Examples	Process component
Role/function of emotion	Attention shift warning, performance evaluation, malfunctioning-component warning, motivation intensifier	Emotion-signal
Process by which emotion fulfills its role	Reflexive action, creation of motivators, setting plan selection criteria	Emotion-response
Urgency of the repairing process	Urgent (e.g. need to immediately attend the external environment), not urgent (e.g. need for long-term improvement of default criteria for plan selection)	Emotion-response
Source of appraisal	External environment, internal state, past events, current events	Appraisal stage
Type of appraisal	Affective appraisal, cognitive appraisal	Appraisal stage

Table 10: Dimensions of emotion classification (according to Botelho and Coelho, 1997, p. 5)

The authors distinguish between affective and cognitive assessment and claim that it is in principle possible to clearly distinguish between these two components in a given architecture. They refer to the corresponding modules as *Affective Engine* and *Cognitive Engine*.

The *Affective Engine* and the *Cognitive Engine* differ in three aspects:

a. Type of information processed: The *affective engine* processes information that has to do with the hypothetical or actual satisfiability of the agent's motives, while the *cognitive engine* also processes problem-solving information, decision information, and declarative information about different aspects of the world.
b. Goal of information processing: The main goal of the information processing of the *affective engine* is to generate signals that help the *cognitive engine* to perform its tasks, for example, selecting the cognitive structures relevant to a situation, controlling attention, etc. The main goals of the *Cognitive Engine* are goal achievement, problem solving and decision-making. "A simple way to put it is to say the Cognitive Engine reasons at the object level, whereas the Affective Engine reasons at the meta-level." (Botelho and Coelho, 1997, p. 10).

c. Typical response time: The *affective engine* reacts much faster than the *cognitive engine* because it requires only a fraction of the information and its architecture also contributes to faster decisions.

The authors propose a mechanism that gives the *affective engine* the ability to react quickly: the reduction of explicit and long chains of comparison into short, specific rules. They give an example of such a process:

if someone risks dying, he or she will feel a lot of fear;
risks_dying(A) -> **activate**(fear, negative, 15)
if someone risks running out of food, he or she risks dying;
risks_running_out_of_food(A) -> risks_dying(A)
if someone risks running out of money, he or she risks running out of food;
risks_running_out_of_money(A) -> risks_running_out_of_food(A)
if someone loses some amount of money, he or she risks running out of money;
loses_money(A) -> risks_running_out_of-money(A)

loses_money(A) -> **activate**(fear, negative,15)

(Botelho und Coelho, 1997, S. 11)

These explicit and implicit rules should be organized in a hierarchy where the longer rules are only used if no short matching rule is found.

The *Salt & Pepper* architecture consists of three main components: the *Affective Engine*, the *Cognitive Engine* and an *Interrupt Manager*. The *Affective Engine* has *Affective Sensors*, an *Affective Generator* and an *Affective Monitor*. The last two initiate the process of generating emotions together. All other modules of the system (except the *Interrupt Manager*) are *to be attributed to* the Cognitive Engine.

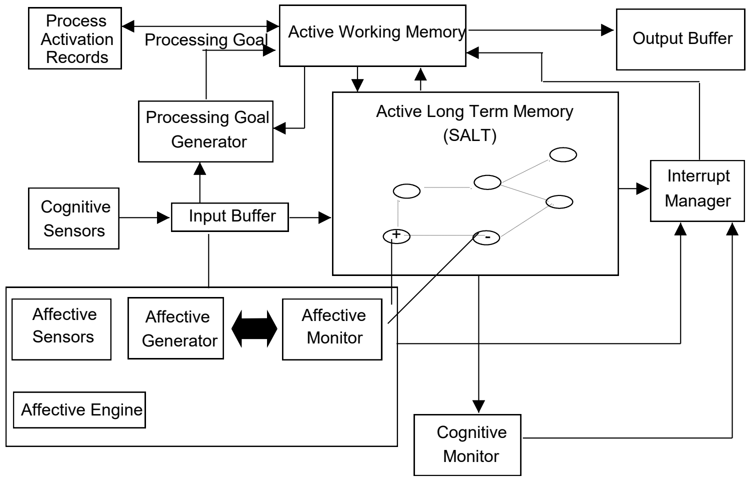


Fig. 21: *Salt & Pepper architecture* (after Botelho and Coelho, 1997, p. 12)

Long-term memory is an associative network. Each node of the network possesses an identification, an activation level, a set of associations with other nodes, and a set of symbolic structures that represent motives, plans, actions, and declarative knowledge. The more a node is activated, the more likely it is to be noticed in a search process (*accessability*).

The *Input Buffer* and the *Affective Generator* activate nodes in long-term memory. The *Cognitive Monitor* and the *Affective Monitor* suggest specific nodes for the agent's attention. Whenever such a suggestion process is running, the *interrupt manager* decides whether to interrupt the currently running cognitive process and load the contents of the proposed node into memory to be processed. When the contents of a node are processed in memory, the node receives a certain level of activation and thus more *accessibility*.

Nodes that are based on certain experiences of the agent are called episodic nodes and form episodic memory.

Emotions are described in this system by a number of parameters:

a. a label E, which describes the emotion class and a list of arguments, for example, the source of the assessment;
b. a valence V, which can assume the values positive, negative or neutral;
c. an intensity I;
d. an emotion program P, which represents a sequence of actions that is executed as soon as the assessment level has produced a label;
e. an emotional response R, which is only executed when a node in the long-term memory that matches the label of the emotion is selected and processed in memory.

The emotion program differs from the emotional response in that it is executed by the *Affective Generator* without interrupting the agent's ongoing cognitive processing.

The *affective generator* performs a partial evaluation of the agent's external and internal state, the so-called affective assessment. When the invocation conditions of a particular emotion are met, the *Affective Generator* generates the label, intensity, and valence of the emotion and executes the emotion program. The *Affective Monitor* then searches the long-term memory until it finds a node that corresponds to the label of the emotion and has the same valence. He activates it with an activation level that is a function of the generated emotional intensity.

The system has mechanisms that allow it to learn emotionally. The authors distinguish between three classes of emotional learning:

1. Learning new and optimizing assessment rules: This includes learning new circumstances that can trigger an emotion signal, reducing assessment rules (see above) and changing the characteristics of the generated emotion signal.
2. Expansion of the repertoire of emotional signals.
3. Learning emotional responses: This includes optimizing existing behavioral responses, learning new responses, and learning as a result of responding to an emotion signal.

The authors specify the conditions under which a system is able to carry out these learning processes (Botelho and Coelho, 1998).

Some elements of *Salt & Pepper* have been implemented so far and, according to the authors, have confirmed the theoretical assumptions (Botelho and Coelho, 1997).

.6. The model of Canamero

Canamero (1997) also follows an approach based on Minsky's "Society of Mind" (1985). In a two-dimensional world called *Gridland* live the *Abbotts*, artificial creatures that also have a motivational and emotional system.

An *Abbott* consists of a large number of agents who, each taken individually, are "simple", but when combined they reach a new quality. An *Abbott* has three types of sensors (somatic, tactile, visual); two types of *recognizers* that respond to complex stimuli and can both learn and forget; eight so-called *direction-nemes*, which provide information from the *Abbott'*s spatial environment ; two categories of *maps* (tactile and visual) that receive their information from the *recognizers,* and *direction-nemes* and represent them internally; three effectors (hand, foot, mouth); a behavioral repertoire (*attack, drink, eat, play, rest, withdraw,* etc.) and a number of *managers* (e.g. *finder*, *look-for*, *go-toward*) that correspond to appetizing behavior. In addition, *the Abbotts* have a number of physiological variables, e.g. adrenaline, blood sugar, endorphins, body temperature, etc.

The *Abbotts* move in a world where there are food sources and obstacles, as well as enemies. They come into this world as "newborns" with a basic set of characteristics and then have to develop in their environment.

What is interesting about the Canamero model is that its creatures are endowed with motivations and emotions from the very beginning. They are called, according to Minsky, *proto-specialists* because they are primitive mechanisms that are responsible for action selection and control functions.

The theoretical basis for the motivations is a homeostatic approach:

> "In general, motivations can be seen as homeostatic processes which maintain a controlled physiological variable within a certain range."
> (Canamero, 1997, S. 6)

The Abbotts' *motivational agents* consist of

> "a controlled variable, the set point and the normal variability range of which are defined by the corresponding sensor that tracks ist real value; an incentive stimulus that can increase the motivation's activation level, but cannot trigger it; an error signal or drive; and a satiation criterion."
> (Canamero, 1997, S. 6f.)

For example, the error message "too low blood sugar" invokes the motivation hunger, whose goal is to increase blood sugar levels. The activation of a motivation is proportional to the size of the error message (or deviation of a physiological value from the homeostatic state); the intensity of motivation is also calculated based on the level of activation. The motivation with the highest level of activation tries to organize the behavior of the *Abbott* in such a way that the corresponding drive is satisfied. If the motivation can't find and invoke appropriate behavior, it activates the *finder agent* and passes the intensity value to it so that it can pass it on to other agents that it activates. The intensity has a significant influence on a behavior: in the case of flight behavior, for example, the strength of the motor activity, in other behaviors, for example, its duration.

The level of activation and intensity of a motivation can now be modified by emotions. In Canamero's system, emotions consist of

> "an incentive stimulus; an intensity proportional to its level of activation; a list of hormones it releases when activated; a list of physiological symptoms; and a list of physiological variables it can affect."
> (Canamero, 1997, S. 7)

Emotional states are activated and distinguished from each other by three types of triggers:

a. External events, i.e. an object or the result of a behavior, where the reaction to it can be either innate or learned.
b. General stimulation patterns that cause different changes in the physiological variables and thus allow the same emotion to act in different circumstances. As an example, Canamero mentions the *anger agent*, which is called by a persistently too high level of a variable. As a result, emotions contribute to the control of homeostatic processes.
c. Special value patterns of physiological variables that allow a distinction to be made between emotions that are evoked by the same general mechanism. As examples, Canamero cites fear (with a high heartbeat rate) and interest (with a low heartbeat rate).

Since *Abbott* is a primitive system, it is always in a clear emotional state. The three triggers are hierarchically arranged in the order of the above list. The selected emotion influences the action selection mechanism in two ways: it can reduce or increase the intensity of the current motivation and thus also the intensity of the selected behavior; it also modifies the results of the sensors that measure the variables that influence emotion, thus altering the perceived physical state (*happiness agent* -> release of endorphin -> lower perception of pain).

The action selection of an *Abbott* thus takes place in four stages:

1. The activation level of all agents is set to zero.
2. Internal variables and environmental data are read in and *maps* are formed.
3. Motivations are assessed and the effects of the emotional state are calculated. The motivation with the highest activation is selected.
4. Active motivation selects the behavior(s) that can best satisfy its instinct.

Canamero acknowledges that their *Abbotts* are currently operating at a very primitive level and need a number of additional agents to develop long-term learning and strategies. Emotions play an essential role in her model:

> "In particular, as far as learning is concerned, our model of emotions provides a means to have different reward and punishment mechanisms... Again, motivations and emotions constitute a key factor in determining what has to be remembered and why."
> (Canamero, 1997, S. 8)

11.7. Summary and Evaluation

As the previous examples show, there is now a variety of approaches to modeling emotions in the field of autonomous agents. The references to psychological theory are varied.

It is striking that most authors proceed quite eclectically in their theoretical borrowings. Mainly the theories that are well suited for operationalization are used. Often only certain elements are singled out, which are then expanded by their own components, often without this being made explicitly clear.

In order to achieve fast results, only partial areas of the designed models are implemented in real simulations or robots. Pragmatic solutions are used that necessarily reduce complex processes to a few variables. In addition, these variables are often defined arbitrarily in order to be able to implement the model at all.

It is striking that the majority of the authors consider emotions to be an essential part of an agent's control system and define them functionally in this respect. Emotions are no longer seen as appendages of the cognitive system, but rather as an indispensable prerequisite for the reliable functioning of cognition.

However, a look at the models also shows that all of them are still far from being able to implement the postulated claim in reality.

12. Significance for Emotion Psychology Research

The computer models of emotions presented in this thesis were created at different times, under different technical conditions and sometimes under completely different initial conditions. Nevertheless, they have one thing in common: the conviction that the modeling of emotions in the computer will contribute to the progress of knowledge.

Emotion psychology research has so far kept a low profile in this area. One can only speculate about the reasons for this. Perhaps it is ignorance of developments in this area (which this thesis would like to make a small contribution to remedying); or perhaps it is the (unconscious) conviction that emotions are a profoundly human phenomenon (and perhaps that of some animal species) that a machine cannot dispose of.

So the field was and is left to computer scientists and roboticists, whose basis is of course not primarily psychological theory. It is then easy to criticize their approaches and point out their lack of psychological foundation. This is a simple way, but also a wrong one.

Michael Gazzaniga, one of the world's leading neuroscientists, writes in the introduction to his latest book:

> “Today, the mind sciences are the province of evolutionary biologists, cognitive scientists, neuroscientists, psychophysicists, linguists, computer scientists - you name it. (...) Psychology itself is dead. (...) The odd thing is that everyone but its practitioners knows about the death of psychology.”
> (Gazzaniga, 1998, S. 11f.)

One of the reasons for such a judgment is that many psychologists have for too long refused to work constructively with other disciplines. In doing so, they have left central questions of psychology, for example that of "consciousness", to others. It was only after philosophers, computer scientists and neuroscientists had already advanced the current discussion on this topic that psychologists began to participate.

A similar fate could threaten emotion psychology. While the psychologists argue about questions such as: What emotions are there? Are there basic and derived emotions? etc., computer scientists, neurologists, biologists and philosophers are concerned with the evolutionary function of emotions, with the emergence of emotions in the brain and the relationship between emotions and consciousness. Here, too, many essential questions are asked outside of psychology - and emotion psychology research largely ignores them.

The computer modeling of emotions includes, as the present work shows, much more than "just" the implementation of a psychological theory in a computer model. It raises a number of fundamental problems relating to the role of emotions in a complex overall system.

Computer models of emotions are therefore interesting for emotion psychology research from several points of view, of which testing an emotion theory against a functioning model is only the most obvious.

However, even the implementation of a psychological theory in a computer model raises problems that have nothing to do with the original theory. One group of problems has to do with the often poor performance of the available computers. Frijda explains this using the example of ACRES, in which some elements of the underlying theory could not be implemented; Sloman also has to resort to an artifice in MINDER1 in order to be able to represent the essential element of his theory in the model at all.

The second group of problems are additions or deviations from the underlying theory, which result from pragmatic solutions in the programming of the model, without being addressed as components of the theory. Chwelos and Oatley make this clear in their critique of Scherer's GENESIS. Elliott has also extended the theory of Ortony, Clore and Collins for his Affective Reasoner by two categories of emotions, without making clear what this extension means for the underlying theory.

Finally, Reilly takes a much more radical approach by replacing an entire part of the theory of Ortony, Clore and Collins with his own construct, which is a kind of cognitive "shortcut" in the recognition of emotions. Of course, he massively changes the basic assumptions of the underlying theory. However, he does not address the impact of these changes on the theory of Ortony, Clore and Collins.

The models of Elliott and Reilly cannot, strictly speaking, be regarded as confirmations of the theory of Ortony, Clore and Collins, since neither of them adopts the theoretical foundations unchanged.

Now, Elliott and Reilly's interest is not in testing a psychological theory of emotion. They only fall back on the emotional model of Ortony, Clore and Collins because it is easy to operationalize. But it is precisely this that makes her finding that the model cannot be implemented one-to-one all the more interesting. Unfortunately, a reaction of Ortony, Clore and Collins to these changes to their model and their significance for their theory is not known.

Computer models that are supposed to test a psychological theory of emotion come from Scherer and Frijda. In particular, Frijda's approach demonstrates very nicely the interaction between theory and model: After developing his first model ACRES, Frijda recognized some shortcomings of his original theory, which he then modified. On the basis of this modified theory, the computer model WILL is currently being developed, the implementation of which will lead to new insights that will either confirm the theory or make its further modification clear.

In addition to testing theories of emotion, computer models are also of interest for emotion psychology research from other points of view. They can help to gain insights into both the function and the functional mode of action of emotions.

There has been much speculation about the function of emotions. Explanations range from the role of a reaction system essential for survival without cognitive components to a subsystem without which rational decisions cannot be made. Computer models can help to provide more clarity here.

For example, both Oatley and Johnson-Laird as well as Sloman postulate that emotions represent a control system that is indispensable for an intelligent system. Such a theory can only be empirically tested on the basis of a computer model (if it can currently be tested at all).

In addition to the function of emotions, the functional mode of action of emotions is also the subject of a number of speculations. It seems undisputed that a hedonistic component plays a role in this. Even theoretical assumptions about such a mode of action can currently only be tested using a computer model, as Wright is trying to do, for example.

Pfeifer and a number of other researchers criticize the computer models of emotions outlined for ignoring an essential factor: physicality. For them, emotions (and cognition in general) are inextricably linked to a body. The most radical approach is to let robots go through an evolutionary process in order to be able to observe the emergence of emotions. These robots are not given any predefined or pre-programmed emotions from the outset.

It is indisputable that the results of such an approach will certainly also bring a number of important insights for research in the psychology of emotions, because in the best case it will actually be possible to follow the emergence of an emotional subsystem in the evolutionary process.

The success of cognitive science is due in no small part to the development of functional models of how the human mind works, which were then implemented in computer models. A prominent example is the recognition and processing of visual stimuli. By concentrating on the functional aspect of psychological processes, it was possible to abstract from the specific hardware (or wetware) and to achieve results with the help of computers that could not have been achieved (or only much later) without this possibility of implementation.

In psychological theory, too, the conception of emotions as functional phenomena is gaining more and more space. Thus, they also represent an area of research to which computer models can make an important contribution.

Of course, different theories of emotion have a different degree of operationalizability. However, most current models belong to the group of estimation theories and thus do not in principle escape computer modelling. In addition, many processes related to the emergence of an emotion elude introspection and thus one of the most popular tools of emotion psychology research. We are aware that we hear something; but we are not aware of how our brain converts the sound waves hitting the eardrum into something that we can perceive as meaningful sound. Likewise, we are often aware that we are in an emotional state, but not how our brain and body created that state. And above all: We cannot consciously switch off this state, just as a tinnitus sufferer cannot stop the whistling in his ear by a deliberate act.

The development of theories about these processes that are not accessible to introspection and their testing with the help of computer models can therefore be of great benefit to emotion psychology research - and thus perhaps also to people who suffer from certain emotional phenomena.

The fact that many impulses in computer modeling of emotions come from areas such as philosophy, artificial intelligence, robotics or neuroscience offers an opportunity to approach the topic in an interdisciplinary manner and thus overcome the horizon of one's own field and its necessary limitations. It doesn't help if one person points the finger at the other and makes fun of his ignorance of the history of emotion psychology theory or machine learning algorithms. An interdisciplinary study of emotions can only be useful for the progress of knowledge in this field without depriving each individual subject area of its very own competence.

Randolph Cornelius writes at the end of his excellent introductory book to the psychology of emotions: "Perhaps there is hope for a harmonic convergence after all." (Cornelius, 1996, p. 213) He relates this remark to the convergence between the four different theoretical approaches to emotions in emotion psychology, which he has noted to some extent. Perhaps this hope also exists with regard to a convergence between the different fields of knowledge that deal with emotions today.

Literature

Affective Computing Home Page unter *http://www-white.media.mit.edu/vismod/demos/affect/affect.html*.

Araujo, A.F.R. (1994). *Memory, Emotions, and Neural Networks: Associative Learning and Memeory Recall Influenced by Affective Evaluation and Task Difficulty*. PhD thesis, University of Sussex.

Aubé, M. (1998). *Designing Adaptive Cooperating Animats Will Require Designing Emotions: Expanding upon Toda's Urge Theory*. Paper presented at the 5th International Conference of the Society for Adaptive Behavior, Zürich. Erhältlich unter http://www.ai.univie.ac.at/~paolo/conf/sab98/sab98ws.html.

Axelrod, R. (1990). *The Evolution of Co-operation*. Penguin Books.

Bareiss, R. (1989). *Exemplar-Based Knowledge Acquisition. A Unified Approach to Concept Representation, Classification, and Learning*. Academic Press.

Bates, J. (1994). *The role of Emotion in Believable Agents*. Communications of the ACM.

Bates, J.; Loyall, A. B.; Reilly, W. S. (1992a). *An Architecture for Action, Emotion, and Social Behavior*. In: *Proceedings of the Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agent World*. S. Martino al Cimino, Italien.

Bates, J.; Loyall, A. B.; Reilly, W. S. (1992b). *Integrating Reactivity, Goals, and Emotion in a Broad Agent*. In: *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Bloomington, Indiana.

Baumgartner, P. und Payr, S. (eds). (1995). *Speaking Mindes. Interviews with Twenty Eminent Cognitive Scientists*. Princeton University Press.

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

Botelho, L.M. und Coelho, H. (1997). *Artificial autonomous agents with artificial emotions*. Erhältlich unter http://iscte.iscte.pt/~luis/web/luis.htm

Botelho, L.M. und Coelho, H. (1998). *Adaptive agents: emotion learning*. Paper presented at the 5th International Conference of the Society for Adaptive Behavior, Zürich. Erhältlich unter http://www.ai.univie.ac.at/~paolo/conf/sab98/sab98ws.html.

Braitenberg, V. (1993). *Vehicle. Experiments with cybernetic beings*. Reinbek, Rowohlt.

Brustoloni, J. C. (1991). *Autonomous Agents: Characterization and Requirements*. Carnegie Mellon Technical Report CMU-CS-91-204. Pittsburgh: Carnegie Mellon University.

Canamero, D. (1997). *Modelling motivations and emotions as a basis for intelligent behavior*. In: *Proceedings of Agents '97*. ACM.

Chwelos, G. und Oatley, K. (1994). *Appraisal, Computational Models, and Scherer's Expert System.* In: *Cognition and Emotion*, 8 (3), S. 245-257.

Colby, K.M. (1981). *Modeling a paranoid mind*. In: *The Behavioral and Brain Sciences*, 4 (4), S. 515-560.

Cornelius, R.R.(1996). *The Science of Emotion*. Prentice Hall.

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

Dawkins, M.S. (1993). *Through Our Eyes Only? The Search for Animal Cosciousness*. Oxford, W.H. Freeman.

Dawkins, R. (1988). *The Blind Watchmaker*. Penguin Books.

Dennett, D. (1996). *Kinds of minds. Toward an Understanding of Consciousness*. Basic Books.

Dörner, D. und Hille, K. (1995). *Artificial Souls: Motivated Emotional Robots*. Erhältlich unter http://www.uni-bamberg.de/~ba2dp1/psi.htm.

Dörner, D., Hamm, A., Hille, K. (1997). *EmoRegul. Description of a program for simulating the interaction of motivation, emotion and cognition in action regulation.* Available at http://www.uni-bamberg.de/~ba2dp1/psi.htm

Dörner, D. and Schaub, H. (1998). *The life of PSI. On the interplay of cognition, emotion and motivation - or: A simple theory for complicated behaviors.* Available at http://www.uni-bamberg.de/~ba2dp1/psi.htm.

Dyer, M.G. (1982). *In-depth understanding. A computer model of integrated processing for narrative comprehension.* Cambridge, Mass., MIT Press.

Dyer, M.G. (1987). *Emotions and their Computations: Three Computer Models*. In: *Cognition and Emotion*, 1 (3), S. 323-347.

Elliott, C. (1992). *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System.* Ph.D. Dissertation, Northwestern University, The Institute for the Learning Sciences, Technical Report No.32.

Elliott, C. (1994a). *Research problems in the use of a shallow artificial intelligence model of personality and emotion.* In: *Proceedings of the Twelfth National Conference on Artificial Intelligence.* Seattle, WA: AAAI.

Elliott, C. (1994b). *Components of two-way emotion communciation between humans and computers using a broad, rudimentary, model of affect and personality.* In: *COGNITIVE STUDIES: Bulletin of the Japanese Cognitive Science Society,* 1(2):16-30.

Elliott, C. (1997a). *I picked up catapia and other stories: A multimodal approach to expressivity for "emotionally intelligent" agents.* In: *Proceedings of the First International Conference on Autonomous Agents*.

Elliott, C. (1997b). *Hunting for the Holy Grail with "emotionally intelligent" virtual actors.* http://condor.depaul.edu/~elliott.

Elliott, C., and Siegle, G. (1993). *Variables influencing the intensity of simulated affective states.* In: *AAAI technical report for the Spring Symposium on Reasoning about Mental States: Formal Theories and Applications*. American Association for Articial Intelligence, Stanford University, Palo Alto, CA..

Elliott, C.; Yang, Y.-Y.; Nerheim-Wolfe, R. (1993). *Using faces to express simulated emotions.* Unpublished manuscript.

Elliott, C., and Carlino, E. (1994). *Detecting user emotion in a speech-driven interface.* Work in progress.

Elliott, C.; Rickel, J.; Lester, J.C. (1997). *Integrating affective computing into animated tutoring agents.* Erhältlich unter http://condor.depaul.edu/~elliott.

Foliot, G. und Michel, O. (1998). *Learning Object Significance with an Emotion based Process*. Paper presented at the 5th International Conference of the Society for Adaptive Behavior, Zürich. Erhältlich unter http://www.ai.univie.ac.at/~paolo/conf/sab98/sab98ws.html.

Franklin, S. (1995). *Artificial Minds*. Cambridge, MA, MIT Press.

Franklin, S. und Graesser, A. (1996). *Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents.* In: *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*. Springer Verlag.

Frijda, N.H. (1986). *The emotions*. Cambridge, U.K., Cambridge University Press.

Frijda, N.H. und Swagerman, J. (1987). *Can computers feel? Theory and design of an emotional system.* In: *Cognition and Emotion*, 1, S. 235-258.

Frijda, N.H. und Moffat, D. (1993). *A model of emotions and emotion communication*. In: *Proceedings of RO-MAN '93: 2nd IEEE International Workshop on Robot and Human Communication*.

Frijda, N.H. und Moffat, D. (1994). *Modeling emotion*. In: *Cognitive Studies*, 1:2, S. 5-15.

Gadanho, S.C. und Hallam, J. (1998). *Emotion-triggered Learning for Autonomous Robots*. Paper presented at the 5th International Conference of the Society for Adaptive Behavior, Zürich. Erhältlich unter http://www.ai.univie.ac.at/~paolo/conf/sab98/sab98ws.html.

Gazzaniga, M.S. (1998). *The mind's past.* Berkeley, University of California Press.

Holland, J.H. (1995). *Hidden Order. How adaptation builds complexity*. Reading, MA, Helix Books.
Holland, J.H. (1998). *Emergence. From Chaos to Order*. Reading, MA, Helix Books.
Kaiser, S. und Wehrle, T. (1993). *Emotion research and AI: Some theoretical and technical issues.* Erhältlich unter http://www.unige.ch/fapse/emotion/.

Kaiser, S. et. al. (1994). *Multi-modal emotion measurements in an interactive computer-game: A pilot-study*. In: N. Frijda (ed.): *Proceedings of the VIIIth Conference of the International Society for Research on Emotion,* 1994.

Koda, T. (1997). *Agents with Faces: A Study on the Effects of Personification of Software Agents.* Thesis, MIT.

LeDoux, J. (1996). *The Emotional Brain*. Simon & Schuster.

McCarthy, John (1990). *Formalizing Common Sense.* Norwood, Ablex Publishing Corporation.

McFarland, D. und Bösser, T. (1993). *Intelligent behavior in animals and robots.* Cambridge, MA, MIT Press.

Minsky, M. (1987). *Societies of Mind.* Picador.

Moffat, D. (1997). *Personality parameters and programs*. In: R. Trappl und P. Peta (eds.): *Creating personalities for synthetic actors*. Springer.

Moffat, D., Frijda, N.H., Phaf, H. (1993). *Analysis of a model of emotions*. In: A. Sloman, D. Hogg, G. Humphreys, A. Ramsay, D. Partridge (eds.): *Prospects for Artificial Intelligence*. Amsterdam, IOS Press.

Moffat, D. und Frijda, N.H. (1995). *Where there's a Will there's an Agent*. In: M.J. Woolridge und N.R. Jennings (eds.): *Intelligent Agents - Proceedings of the 1994 Workshop on Agent Theories, Architectures, and Languages*. Springer.

Mueller, E. und Dyer, M.G. (1985). *Daydreaming in humans and computers*. In: *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*. Los Angeles, CA.

Neisser, U. (1963). *The imitation of man by machine*. In: *Science*, 139, S. 193-197.

Oatley, K. (1992). *Best Laid Schemes. The Psychology of Emotions*. Paris, Cambridge University Press.

Oatley, K. und Johnson-Laird, P.N. (1987). *Towards a cognitive theory of emotions.* Cognition and Emotion, 1, 29-50.

Oatley, K. und Jenkins, J.M. (1996). *Understanding Emotions.* Blackwell.

O'Rorke, P. und Ortony, A. (1992). *Explaining emotions*. Unveröffentlichtes Manuskript.

Ortony, A., Clore, G.L., Collins, A. (1988). *The cognitive structure of emotions.* Cambridge, U.K., Cambridge University Press.

Pfeifer, R. (1982). *Cognition and emotion: an information processing approach.* Carnegie-Mellon University, CIP Working Paper Nb. 436.

Pfeifer, R. (1988). *Artificial intelligence models of emotion.* In: V. Hamilton, G. Bower, & N. Frijda (eds.). *Cognitive perspectives on emotion and motivation. Proceedings of the NATO Advanced Research Workshop*. Dordrecht, Kluwer.

Pfeifer, R. (1994). *The "Fungus Eater" approach to the study of emotion: A View from Artificial Intelligence.* Techreport #95.04. Artificial Intelligence Laboratory, University of Zürich.

Pfeifer, R. (1996). *Building "Fungus Eaters": Design Priciples of Autonomous Agents.* In: *Proceedings of the Fourth International Conference of the Society for Adaptive Behavior.* Cambridge, MA, MIT Press.

Pfeifer, R. (1998). *Cognition*. In L. Steels (ed.). *The biology and technology of intelligent autonomous agents. Proceedings of the NATO Advanced Study Institute*. Trento, Italien.

Pfeifer, R. (1998). *Cheap designs: exploiting the dynamics of the system-environment interaction.* Technical Report No. IFI-AI-94.01, AI Lab, Computer Science Department, University of Zurich.

Pfeifer, R. und Nicholas, D.W. (1985). *Toward computational models of emotion.* In: L. Steels, and J.A. Campbell (eds.).: *Progress in Artificial Intelligence.* Chichester, U.K., Ellis Horwood.

Pfeifer, R. und Verschure, P.F.M.J. (1992). *Distributed adaptive control: a paradigm for designing autonomous agents.* In: *Toward A Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life.* Cambridge, MA, MIT Press.

Pfeifer, R. und Verschure, P.F.M.J. (1998). *Complete autonomous agents: a research strategy for cognitive science.* In: G. Dorffner (ed.): *Neural networks and a New AI.*

Picard, R.W.(1997). *Affective Computing.* MIT Press, Cambridge MA.

Read, T. und Sloman, A. (1993). *The Terminological Pitfalls of Studying Emotion*.

Reeves, J.F. (1991). *Computational morality: A process model of belief conflict and resolution for story understanding*. Technical Report UCLA-AI-91-05, UCLA Artificial Intelligence Laboratory.

Reilly, W.S. (1996). *Believable Social and Emotional Agents*. PhD thesis. Technical Report CMU-CS-96-138, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Reilly, W. S. und Bates, J. (1992). *Building Emotional Agents.* Technical Report CMU-CS-92-143, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Rollenhagen, C. und Dalkvist, J. (1989). *Cognitive contents in emotion: A content analysis of retrospective reports of emotional situations*. Technical Report, Department of Psychology, University of Stockholm.

Roseman, I.J. (1979). *Cognitive aspects of emotion and emotional behavior*. Paper presented at the 87th Annual Convention, American Psychological Association. New York, NY.

Roseman, I.J. (1984). *Cognitive determinants of emotions: A structural theory*. In: P. Shaver (Ed.): *Review of personality and social psychology, Vol. 5*. Beverly Hills, CA, Sage.

Roseman, I.J. (1991). *Appraisal determinants of discrete emotions.* In: *Cognition and Emotion*, 3, 161-200.

Roseman, I.J; Antoniou, A.A.; Jose, P.A. (1996). *Appraisal Determinants of Emotions: Constructing a More Accurate and Comprehensive Theory*. In: *Cognition and Emotion*, 10 (3), S. 241-277.

Schaub, H. (1995). *Die Rolle der Emotionen bei der Modellierung kognitiver Prozesse.* Paper zum Workshop Artificial Life, Sankt Augustin. Erhältlich unter http://www.uni-bamberg.de/~ba2dp1/psi.htm.

Schaub, H. (1996). *Künstliche Seelen - Die Modellierung psychischer Prozesse.* Widerspruch 29.

Scherer, K. (1984). *On the nature and function of emotion: a component process approach.* In K.R. Scherer, and P. Ekman (eds.). *Approaches to emotion.* Hillsdale, N.J., Erlbaum.

Scherer, K. (1988). *Criteria for emotion-antecedent appraisal: A review*. In: V. Hamilton, G.H. Bower, N.H. Frijda (eds.): *Cognitive perspectives on emotion and motivation*. Dordrecht, Kluwer.

Scherer, K. (1993). *Studying the Emotion-Antecedent Appraisal Process: An Expert System Approach*. In: *Cognition and Emotion*, 7 (3/4), S. 325-355.

Selfridge, O.G.(1959). *Pandemonium: A Paradigm for Learning.* In: Blake, D.V. and Uttley, A.M. (eds.): *Proceedings of the Symposium on Mechanization of Thought Processes.* H.M. Stationary Office, London.

Simon, H.A. (1967). *Motivational and emotional controls of cognition.* Psychological Review, 74, 29-39.

Simon, H.A. (1996). *Computational Theories of Cognition*. In: W. O'Donohue und R.F. Kitchener (eds.): *The Philosophy of Psychology*. Sage.

Sloman, A. (1981). *Why robots will have emotions.* Proceedings IJCAI.

Sloman, A. (1987). *Motives Mechanisms and Emotions.* In: *Cognition and Emotion* 1,3.

Sloman, A. (1991). *Prolegomena to a theory of communication and affect.* In: A. Ortony, J. Slack, & O. Stock (eds.). *AI and Cognitive Science Perspectives on Communication.* Heidelberg: Springer.

Sloman, A. (1992a). *Towards an information processing theory of emotions.* 1992. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1992b). *Silicon Souls, How to design a functioning mind.* Professorial Inaugural Lecture, Birmingham, 1992. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1992c). *Prolegomena to a Theory of Communication and Affect.* In: Ortony, A., Slack, J., Stock, O. (Eds.): *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues.* Heidelberg, Springer.

Sloman, A. (1993). *Prospects for AI as the General Science of Intelligence*. In: *Proceedings AISB93.*

Sloman, A. (1994). *Explorations in Design Space.* In: *Proceedings 11th European Conference on AI*. Amsterdam.

Sloman, A. (1995). *Exploring design space and niche space.* Invited talk for the 5th Scandinavian Conference on AI, Trondheim, May 1995. In: *Proceedings SCAI95.* IOS Press, Amsterdam.

Sloman, A. (1996a). *What sort of architecture can support emotionality?* Slides for a talk at MIT Media Lab, Nov 1996. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1996b). *What sort of architecture is required for a human-like agent?* Invited talk at Cognitive Modeling Workshop, AAAI96, Portland, Oregon.

Sloman, A. (1997a). *Designing Human-Like Minds.* Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1997b). *Architectural Requirements for Autonomous Human-like Agents.* Slides for a talk at DFKI Saarbrücken, 1997. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1997c). *What sort of control system is able to have a personality?* In: R. Trappl and P. Petta (eds): *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents.* Springer.

Sloman, A. (1998a). *Diagrams in the Mind?* Invited paper for Thinking With Diagrams conference at Aberystwyth. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1998b). *What Sorts of Machines Can Love? Architectural Requirements for Human-like Agents Both Natural and Artificial.* Draft extended version. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1998c). *Review of Affective Computing by Rosalind Picard, MIT Press 1997.* Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1998d). *The ``Semantics'' of Evolution: Trajectories and Trade-offs in Design Space and Niche Space.* Invited talk for IBERAMIA-98 Lissabon.

Sloman, A. (1998e) *Damasio, Descartes, Alarms and Meta-management.* Invited contribution to symposium on Cognitive Agents: Modeling Human Cognition, at IEEE International Conference on Systems, Man, and Cybernetics. San Diego.

Sloman, A.(1998f). *What's an AI toolkit for?* In: B. Logan und J. Baxter (eds.): *Proceedings AAAI-98 Workshop on Software Tools for Developing Agents*.

Sloman, A. (1998g). *Supervenience and Implementation: Virtual and Physical Machines.* Draft. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1998h). *Design Spaces, Niche Spaces and the ``Hard'' Problem.* Draft, Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. (1998i). *What sort of architecture is required for a human-like agent?* In: M Wooldridge and A Rao (eds.): *Foundations of Rational Agency.* Kluwer Academic Publishers.

Sloman, A. und Croucher, M. (1981). *Why robots will have emotions.* In: *Proceedings of the 7th International Joint Conference on AI.* Vancouver.

Sloman, A. und Logan, B. (1997). *Synthetic Minds.* Poster presented at AA'97 Marina del Rey. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. und Logan, B. (1998). *Architectures and Tools for Human-Like Agents.* Paper presented at the European Conference on Cognitive Modelling, Nottingham. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Sloman, A. und Poli, R. (1996). *SIM_AGENT: A toolkit for exploring agent designs.* In: M. Wooldridge, J. Mueller, M. Tambe (eds.): *Intelligent Agents Vol II (ATAL-95).* Springer-Verlag.

Sloman, A. und Wright, I.P. (1997). *MINDER1: An implementation of a protoemotional agent architecture.* Technical Report CSRP-97-1. Erhältlich unter ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/0-INDEX.html.

Smith, C.A., Griner, L.A., Kirby, L.D. Scott, H.S. (1996). *Toward a Process Model of Appraisal in Emotion.* In: *Proceedings of the Ninth Conference of the International Society for Research on Emotions*. Toronto, Canada.

Sonnemans, J. und Frijda, N.H. (1994). *The structure of subjective emotional intensity*. In: *Cognition and Emotion*, 8 (4), S. 329-350.

Staller, A. und Petta, P. (1998). *Towards a Tractable Appraisal-Based Architecture for Situated Cognizers.* Paper presented at the 5th International Conference of the Society for Adaptive Behavior, Zürich. Erhältlich unter http://www.ai.univie.ac.at/~paolo/conf/sab98/sab98ws.html.

Stork, D. G.(1997). *Scientist on the Set: An Interview with Marvin Minsky*. In: Stork D. G. (ed.): *HAL's Legacy: 2001's computer as dream and reality.* MIT Press, Cambridge MA.

Suchman, L. (1987). *Plans and situated actions*. Cambridge University Press.

Sutton, R.S. (1991). *Dyna, an integrated architecture for learning, planning, and reacting*. In: *Working Notes of the 1991 AAAI Spring Symposium*.

Swagerman, J. (1987). *The Artificial Concern Realization System ACRES. A computer model of emotion.* PhD Thesis, University of Amsterdam, Dept. of Psychology.

Toda, M. (1982). *Man, robot, and society*. The Hague, Nijhoff.

Velásquez, J.D. (1997). *Modeling Emotions and Other Motivations in Synthetic Agents.* In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference.* Menlo Park.

Velásquez, J.D. (1998). *A Computational Framework for Emotion-Based Control*. Paper presented at the 5th International Conference of the Society for Adaptive Behavior, Zürich. Erhältlich unter http://www.ai.univie.ac.at/~paolo/conf/sab98/sab98ws.html.

Verschure, P.F.M.J., Kröse, B.J.A., Pfeifer, R. (1992). *Distributed adaptive control: the self-organization of structured behavior.* In: *Robotics and Autonomous Systems*. 9.

Watkins, C. und Dayan, P. (1992). *Technical Note: Q-Learning*. In: *Machine Learning* 8, S. 279-292.

Wehrle, T. (1994). *New fungus eater experiments*. In: P. Gaussier und J.-D. Nicoud (eds.): *From perception to action*. Los Alamitos, IEEE Computer Society Press.

Wilson, S.W. (1995). *Classifier fitness based on accuracy*. In: *Evolutionary Computation*, 3 (2), S. 149-185

Wright, I.P. (1996a). *Design Requirements for a Computational Libidinal Economy*. Technical Report CSRP-96-11. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Wright, I.P. (1996b). *Reinforcement learning and animat emotions.*
Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Wright, I.P. (1997). Emotional Agents. PhD thesis, University of Birmingham. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Wright, I.P. und Sloman, A. (1996). *MINDER1: An Implementation of a Protoemotional Agent Architecture.* Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Wright, I.P. und Aubé, M. (1997). *The society of mind requires an economy of mind.* Technical Report CSRP-97-6. 1997. Erhältlich unter http://www.cs.bham.ac.uk/~axs/cog_affect/COGAFF-PROJECT.html.

Wright, I.P., Sloman, A. und Beaudoin, L. (1996). *Towards a Design-Based Analysis of Emotional Episodes.* In: *Philosophy Psychiatry and Psychology,* vol 3 no 2.