Motivations behind modeling emotional agents: Whose emotion does your robot have?

Thomas Wehrle

FPSE, University of Geneva, Switzerland
9, route de Drize
CH-1227 Carouge-Genève
wehrle@acm.org

Abstract

A communality between research in *Artificial Intelligence* and *Synthetic Emotion* is that it seems in both cases to be rather difficult to give an acceptable definition of the naturally occurring counterpart. One could speculate whether this is due to the multiplicity of the nature of both phenomena or due to a categorical misconception. In this paper I try to briefly outline a number of different motivations for modeling emotions, and to relate those motivations to two different principal design approaches for computational models of emotion. From these two aspects, together with our current assumptions about mechanisms underlying human emotions, I conclude with some speculations about *adaptation* in affective systems, and some implications of the notion of *grounding* emotions in adaptive systems.

Introduction

> It seems certain that, as we understand more about cognition, we will need to explore autonomous systems with limited resources that nevertheless cope successfully with multiple goals, uncertainty about environment, and coordination with other agents. In mammals, these cognitive design problems seem to have been solved, at least in part, by the processes underlying emotions. (Oatley 1987)

This is an example of why one might want to model an emotional agent. I will try to show that there are other motives, and that there are also different principal approaches to model emotions. I will briefly sketch some theory on emotion from a psychological point of view to illustrate our current assumptions about the nature of emotions, and about the structurally and functionally different subsystems that seem to be involved in emotions. Based on these three considerations (motives, modeling approaches, and mechanisms) I would like to make some speculations about possible principles of adaptation in computational models of emotion. I will also try to argue that the notion of *grounding* emotions in adaptive systems raises some interesting questions, and is dependent upon at least the same three aspects.

Different motives for modeling emotional agents

Science (Psychology, Neuroscience, Cognitive Science, Biology, etc.)

Modeling an emotional system can be an attempt to instantiate parts of a theory about a natural phenomenon with a computer program or a robot. The researcher hopes that such a system helps to improve the formalization, operationalisation, and internal consistency of theoretical postulates. The system is further expected to allow an examination of the required number and types of criteria needed for successful theoretical prediction (allowing also to compare theories differing in this respect) and to improve intuitive understanding of these parameters. Such systems can also be very useful in teaching and visualization. This synthetic approach complements the traditional approach of the analysis of behavioral data (and this data can be used to measure the quality of the model).
Criteria: description, explanation, and prediction
Main motivation: Improve our knowledge about the nature of emotion and its implications

Engineering

It seems likely that here the motive behind modeling an emotional agent is an indirect one. The engineer is primarily interested in constructing a useful artifact. In adopting some real or hypothesized natural principles the engineer hopes to increase the system performance in terms of task achievement and costs. The extent to which the principles that inspired the system eventually get into the system or the form and adequacy of the translation is of no significance. Emotion theories may serve as heuristics in finding a good solution to a problem, or to metaphorically describe the state or behavior of a system.
Criterion: Performance
Main motivation: Building good artifacts for a concrete task

Human Computer Interaction

Modeling emotions in a system that interacts with humans is a special case of engineering where human behavior and affectivity plays a significant role. In this case theoretical knowledge about emotions can be applied. In adopting some real or hypothesized natural principles the engineer hopes to increase the system performance in terms of acceptance and usability with respect to the user. Again, it seems likely that such a system does not necessarily need to represent emotion constructs in any form to generate the desired behavior.
Criteria: Performance, acceptance, and usability
Main motivation: Improve human computer interaction

Technology would in this line of reasoning be the attempt to extract general principles from the engineering work.

Two kinds of modeling approaches

Almost since the beginning of the computer age there have been exciting attempts to deal with emotions in one way or another. For examples of good overviews on existing systems I refer to Pfeifer (1988), Picard (1997) and the web page [3] of Hudlicka and Fellous. Given the different concerns of computational models of emotion and affective behavior it is no surprise that we find a whole variety of different modeling approaches. Wehrle and Scherer (1995) have argued that it might be useful to distinguish two classes of computational models of emotion: black box models and process models. Although the two approaches should not been seen as exclusive, they differ in the degree of abstraction of intervening variables[1].

Black box modeling

The purpose of black box models is to produce outcomes or decisions that are maximally similar to those resulting from the operation of naturally occurring systems, disregarding both the processes whereby these outcomes are attained as well as the structures involved (see also Phelps and Musgrove 1986, p. 161).

Although such models provide little information concerning the mechanisms involved, they are very useful for practical decision making and for providing a sound grounding for theoretical and empirical study. In particular, they can help to investigate necessary and sufficient variables. System performance (e.g. quality of classification and computational economy) as well as cost of data gathering are important criteria for assessing the quality of the chosen computational model. Since black box models focus on the input-output relationship, they make few claims whatsoever concerning the nature of the underlying processes.

The simplified example in figure 1 illustrates the idea:

[1] I also think that models that include underlying mechanisms are necessarily based on black box models at a certain level of abstraction.

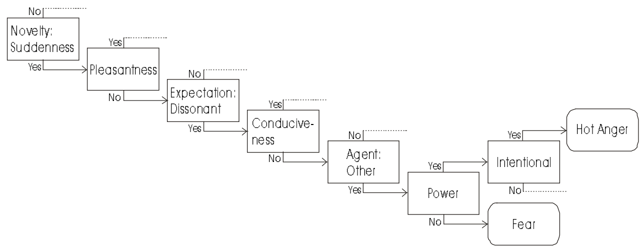

Fig. 1: Part of a decision tree representation of Scherer's Appraisal Theory concerning the cognitive component. Note that the theoretically postulated sequence of appraisal is violated (the intention check should precede the power check) but the black box output is correct.

To my knowledge most existing implementations (whether they simulate or reason about emotions) are based on black box models (and mostly symbol systems).

Process modeling

The purpose of process modeling is usually the attempt to simulate naturally occurring processes using hypothesized underlying mechanisms. Clearly, this approach is considerably more ambitious than the black box approach. In the case of psychobiological process models, one needs to specify the effects of causal input factors on intervening process components (often hypothetical constructs), as well as the structural interdependencies of the internal organismic mechanisms involved.

To my knowledge few systems have attempted to synthesize emotional behavior on this level of biologically plausible mechanisms (e.g. Armony et al. 1995). Given the complexity of involved components, models of this kind seem to be as difficult to realize as they are useful to increase our knowledge. In the Geneva Emotion Research Group we have tried to implement a very simple emotional problem solver inspired by Toda's Social Fungus Eater (Toda 1962, 1982). The system is described elsewhere (Wehrle 1994a 1994b, and on my web page [1]). Toda describes the behavior of the hypothesized Fungus Eaters in terms of urges and cost functions (internal vs. external). In our proposed model we used a simple hedonistic principle based on an energy concept which also allows adaptation. Whereas for Toda the concept of urges seems similar to emotions, we have so far only implemented some very basic urges as value scheme, and regard emotional behavior as an emergent property of the latter.

Possible contributions of appraisal theory

In the following, several theoretical postulates from two prominent theories concerning appraisal and emotion are presented to illustrate the function and conceptualization of emotion in human beings from a psychological point of view. I feel that appraisal theory could potentially offer some heuristics for the design of emotional systems. Since it represents the mainstream of current emotion psychology it may also be used as a source of ideas about how psychologists envisage the complexity of affective behavior. Later on in my argumentation I will also use the assumption that the mechanisms underlying emotions are functionally and structurally heterogeneous to question the notion of grounding emotions in an artifact. Even though the theory is quite general with respect to the emotional situations, it has nevertheless been based upon the specific physical properties of the human organism.

> Scherer has argued that the elicitation and differentiation of emotion can be most economically explained by a process of cognitive event appraisal that reflects the significance of an event for an individual's goal and value structure, his or her coping potential, and the socio-normative significance of the event. The component process theory (see Scherer, 1988, 1993, for details) posits relatively few basic evaluation criteria and assumes a sequential processing of these criteria in the appraisal process. Figure 2 shows the major appraisal dimensions or "stimulus evaluation checks" (SECs) which are considered to be sufficient to account for the differentiation of all major emotions. (Wehrle, Kaiser, Schmidt, and Scherer, submitted)

> Scherer has suggested that emotion can be defined as an episode of temporary synchronization of all major subsystems of organismic functioning represented by five components (cognition, physiological regulation, motivation, motor expression, and monitoring/feeling) in response to the evaluation of an external or internal stimulus event as relevant to central concerns of the organism. It is claimed that while the different subsystems or components operate relatively independently of each other during non- emotional states, dealing with their respective function in overall behavioral regulation, they are recruited to work in unison during emergency situations, the emotion episodes. These require the mobilization of substantial organismic resources in order to allow adaptation or active responses to an important event or change of internal state. The emotion episode begins with the onset of synchronization following a particular stimulus evaluation pattern and ends with the return to independent functioning of the subsystems (although systems may differ in responsivity and processing speed). Since stimulus evaluation is expected to affect each subsystem directly and since all systems are seen to be highly interrelated during the emotion episode, regulation is complex and involves multiple feedback and feedforward processes. For this reason, it is assumed that there is a large number of highly differentiated emotional states, of which the current emotion labels capture only clusters or central tendencies of regularly recurring <u>modal</u> states. … Scherer has suggested that subjective experience or feeling can be conceptualized as the reflection of the changes in all other emotion components, including the different neurophysiological and motor subsystems as well as changes in motivation and particularly the cognitive appraisal system. Leventhal and Scherer (1987) have made a first attempt to illustrate the way in which Scherer's stimulus evaluation checks could be performed on the sensory-motor, the schematic and the conceptual level. Obviously, the nature of the resulting emotion is likely to be quite different depending on the level of its antecedent appraisal, particularly with respect to the conscious experience of the episode. (Kaiser and Scherer 1998)

> Arnold (1960) defined emotions as "felt action tendencies" because, as she argues, a felt tendency is what characterizes such experience and differentiates it from mere feelings of pleasantness or unpleasantness; different action tendencies are what characterize different emotions. Action tendencies or, more generally, changes in action readiness are not only important in emotional experience but are also central in the analysis of emotion as such. Action readiness is what links experience and behavior; felt action readiness can be considered a reflection of the actual state of behavioral readiness…. Emotions involve states of action readiness elicited by events appraised as emotionally relevant… Events are appraised as emotionally relevant when they appear to favor or harm the individual's concerns. (Frijda, Kuipers, and ter Schure)

An attempt to summarize the basic concepts can be found in figure 2. I took some freedom to include some further assumptions about the postulated sequential evaluation and to apply the concept of the different levels of processing also to the expressive component.

Note however, that the theory is based on constructs (for which there is empirical evidence). More directly accessible to us are observable behaviors, some of which I list below:

- Social behavior, problem solving, action selection, etc.
- Facial expression
- Voice and utterances (prosody and syntax)
- Physiological responses (skin conductance, EMG, ECG, EEG, etc.)
- Verbal report of subjective feeling and mood
- Expression in arts

I prefer to leave it open whether this behavior can be seen as an antecedent, as a result, or as a constituent of emotions. My point here is that the great variety of behavior in which emotions play a role seems to reflect the functionally and phylogenetically different mechanisms underlying emotion that we assume.

To further complicate the issue there is evidence that the roles that the different subsystems play, and their connections to the expressive components, vary significantly among different emotions. I refer to the respective literature (concerning psychophysiological measures e.g., Stemmler 1989; and concerning mechanisms, e.g., LeDoux 1996; Panksepp 1993; Davidson 1994).

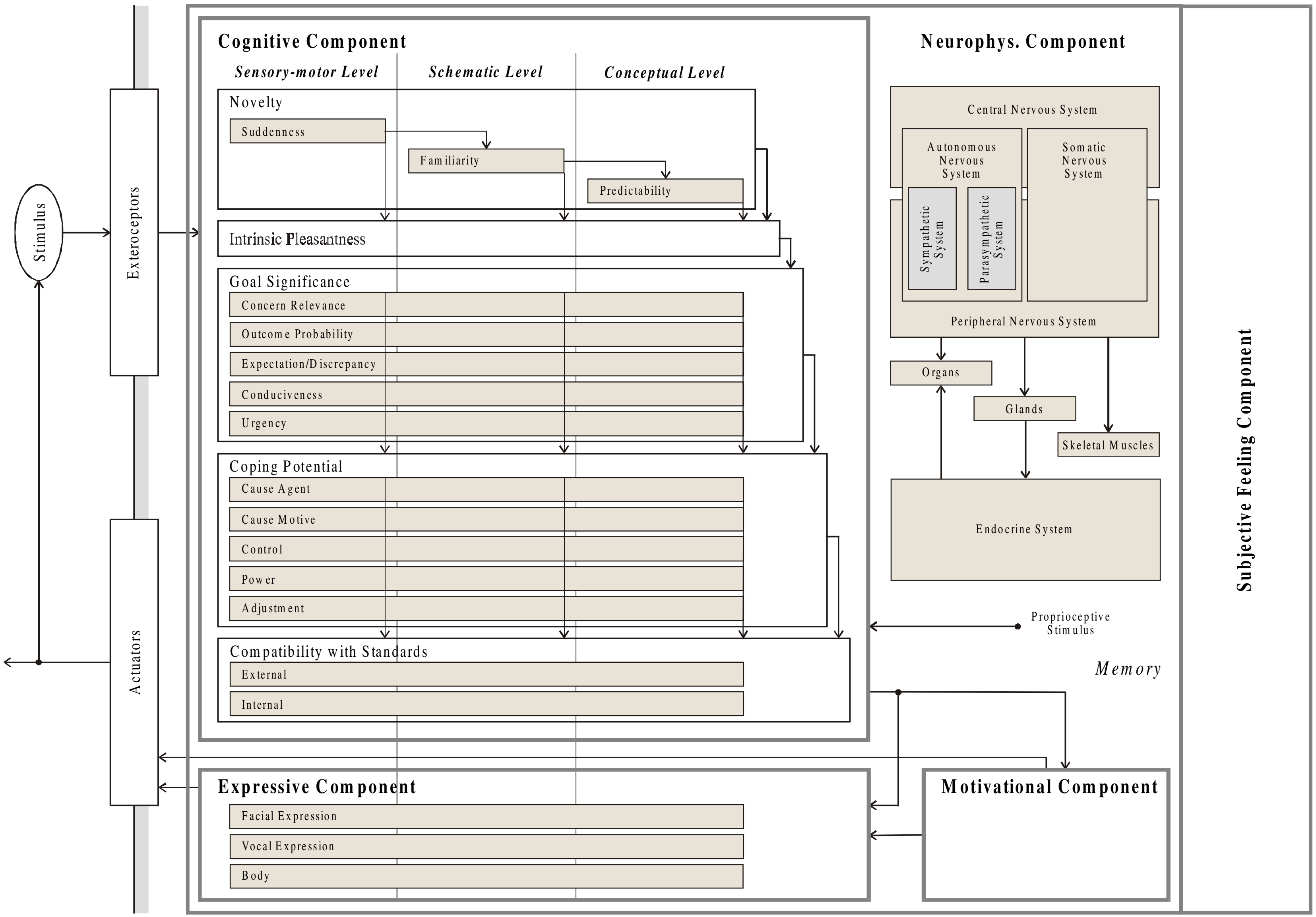

Fig. 2: An attempt of a formal representation of Scherer's Appraisal Theory

Adaptation

If we regard emotions as the result of the interaction and synchronization of rather complex subsystems in response to situational, environmental, and physiological properties, then emotions must reflect an adaptive system. There is evidence that affective behavior is also much more flexible than behavior generated by simpler fixed pattern response systems. There are large individual and cultural differences, and emotional reactions also vary over different life stages. One could speculate about three different levels of processing on which individual adaptation might be implemented:

- Level of single values: One likely candidate where adaptive mechanisms might play a role is in the parameters of the involved stimulus evaluation systems and their influence on the organization of behavior. Example: The attribution of self-competence and power can be derived from experience and changes over time and domain.
- Level of emotion specific value patterns: There seems to be evidence that the significance of different appraisal dimensions varies among emotions. I propose that this is a second candidate for the implementation of adaptive mechanisms. Example: A diplomat might have to learn to reduce the importance of his or her estimated coping potential to avoid aggressing other diplomats in an anger situation.
- Level of action readiness patterns: Since it has been argued that the affective systems allow an organism to generate behavior in an efficient and adapted fashion, the association of an action repertory to an emotional state might also be modified in terms of reflection and success evaluation. Example: In elevators where people are necessarily *physically* closer to each other than they actually would like to be (embarrassment), many people learn to establish a larger *social* distance by actively avoiding eye contact.

Empirical data

Eventually, the quality of a synthetic approach aiming to model human emotion will be compared with the observation of naturally occurring systems. Empirical studies serve also to improve our knowledge about the nature of emotion, the expressive components, and the function of emotion. Most empirical studies rely on verbal report, which is not necessarily the most promising way, since the necessary degree of awareness of an emotional state, and the influence of self-reflection on the emotional process are not yet clear. Isolated studies of certain aspects of emotion such as facial expression on the basis of strong static stimuli seem similarly unpromising. Masanao Toda's proposal of the Fungus Eater experiment, although already described more than 35 years ago, still seems to be a fresh and rarely pursued idea (see also Pfeifer 1994). The Geneva Emotion Research Group is conducting empirical studies within this paradigm with several NSF projects (see [2]; Kaiser and Wehrle, 1996).

Questions and conclusions

I have tried to show that emotions seem to be involved in many functionally different behaviors, and that this impression is also reflected in the many heterogeneous mechanisms that we suppose underlie emotions. I also tried to show that there are different motives behind affective computing, and accordingly different techniques for modeling and representing emotions in an artifact which abstract from the underlying mechanisms to different degrees. I addressed the issue of adaptation by making some speculations about how it could be realized within the framework of appraisal theory.

Unfortunately, it was not possible to fully elaborate the relations between the different aspects of modeling emotions that I described, but I would like to at least conclude with some questions concerning the attempt to *ground* emotion in an *adaptive* system.

As Braitenberg (1984) and others have shown, the observed complexity of behavior does not necessarily need to be reflected in the underlying mechanism. However, if we are to believe that there are different mechanisms underlying emotions, and that they get expressed in functionally different behaviors, the question remains to what extent the supposed structural complexity of involved mechanisms can be abstracted. For certain purposes it has successfully been done in several systems, but the question becomes more important if a system attempts to ground emotions.

If emotion categories are either seen as emergent labels for the evaluation of prototypical situations or events (modal emotions), or as evolutionarily achieved response programs (basic emotions), then we have to expect that emergent emotions in robots will be different than human emotions, i.e. all emotion systems of natural or artificial agents should be expected to be incommensurable if the niches, and probably even more importantly, the physical properties of the agent bodies differ significantly.

Even if we can introduce a type of value system to a robot, I personally feel that grounding somehow implies that we allow the robot to establish its own emotional categorization which refers to its own physical properties, the task, the properties of the environment, and the ongoing interaction with this environment. In this case talking about mechanisms seems unavoidable.

We can put *human* emotion categories into an artificial agent. It might be useful to use emotions as design heuristics for adaptive systems, or to describe their behavior, but can we hope to ground these categories that have evolved in a different system? Is that a reasonable goal? With the framework of appraisal theory I also tried to demonstrate that a theory might be able to abstract from the niche to a certain extent but to a smaller extent from the properties of the agent.

One might argue that implementing a certain value system for an autonomous agent is equivalent to introducing a sort of emotion system in a broad sense. If one takes the appraisal dimensions proposed in emotion psychology as a basis for the value system, one might profit from the possibility to describe the resulting behavior in known emotion terms. Whether or not the chosen values are appropriate depends on the task an agent is designed for, its physical properties, and the properties of the environment. In summary, we can put human emotion categories into an artifact but we will probably not be able to ground them if the properties of the artifact, its tasks, and its environment are significantly different. Or we can let a system develop its own categories, but those might not be the categories that we are used to, and we do not necessarily have to dub these system states with emotion terms.

References

Armony, J. L., Servan-Schreiber, D., Cohen, J. D., and LeDoux, J. E. 1995. An anatomically constrained neural network model of fear conditioning. *Behavioral Neuroscience* 109:246-257.

Braitenberg, V. 1984. *Vehicles: Experiments in synthetic psychology*. Cambridge, Massachusetts: MIT Press.

Davidson, R. J. 1994. Asymmetric brain function, affective style, and psychopathology. The role of early expression and plasticity. *Development and Psychopathology* 6:741-758.

Frijda, N. H., Kuipers, P., and ter Schure, E. 1989. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology* 57:212-228.

Kaiser, S., and Scherer, K. R. 1998. Models of 'normal' emotions applied to facial and vocal expressions in clinical disorders. In *Emotions in psychopathology*, eds. W. F. Flack, Jr., and J. D. Laird, 81-98. New York: Oxford University Press.

Kaiser, S. and Wehrle, T. 1996. Situated emotional problemsolving in interactive computergames. In *Proceedings of the VIIIth Conference of the International Society for Research on Emotions, ISRE'96*. 276-280. Storrs, CT: ISRE Publications.

LeDoux, J. E. 1996. *The emotional brain*. New York: Simon and Schuster.

Oatley, K. 1987. Editorial: Cognitive Science and the understanding of emotions. *Cognition and Emotion* 1:209-216.

Panksepp, J. 1993. Neurochemical control of moods and emotions: From amino acids to neuropeptides. In *Handbook of emotions*, eds. M. Levis and M. J. Haviland, 87-107. New York: The Guilford Press.

Pfeifer, R. 1988. Artificial intelligence models of emotion. In *Cognitive perspectives on emotion and motivation*, eds. V. Hamilton, G. H. Bower, and N. H. Frijda, 287-320. Dordrecht: Kluwer Academic Publishers.

Pfeifer, R. 1994. The 'fungus eater approach' to emotion: A view from artificial intelligence. *Cognitive Studies, The Japanese Society for Cognitive Scienc,* 1:42-57.

Phelps, R. I., and Musgrove, P. B. 1986. Artificial intelligence approaches in statistics. In *Artificial intelligence and statistics*, ed. W. A. Gale. Reading, Massachusetts: Addison-Wesley.

Picard, R. W. 1997. *Affective computing*. Cambridge: The MIT Press.

Scherer, K. R. 1988. Criteria for emotion-antecedent appraisal: A review. In *Cognitive perspectives on emotion and motivation*, eds. V. Hamilton, G. H. Bower, and N. H. Frijda, 89-126. Dordrecht: Kluwer Academic Publishers.

Scherer, K. R. 1993. Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition and Emotion* 7:325-355.

Stemmler, G. 1989. The autonomic differentiation of emotions revisited: Convergent and discriminant validation. *Psychophysiology* 26:617-632.

Toda, M. 1962. Design of a Fungus-Eater. *Behavioral Science* 7:164-183. (Reprinted in Toda 1982, 100-129).

Toda, M. 1982. *Man, robot and society*. The Hague: Martinus Nijhoff Publishing.

Wehrle, T. 1994a. New fungus eater experiments. In *From perception to action,* eds. P. Gaussier and J.-D. Nicoud. Los Alamitos: IEEE Computer Society Press.

Wehrle, T. 1994b. *Eine Methode zur psychologischen Modellierung und Simulation von Autonomen Agenten.* Ph.D. diss, University of Zürich.

Wehrle, T., and Scherer, K. 1995. Potential pitfalls in computational modeling of appraisal processes: A reply to Chwelos and Oatley. *Cognition and Emotion* 9:599-616.

Wehrle, T., Kaiser, S., Schmidt, S., and Scherer, K. R Submitted. Studying dynamic models of facial expression of emotion using synthetic animated faces. *Journal of Personality and Social Psychology*.

WWW links:

[1] Author's home page:
http://www.unige.ch/fapse/emotion/members/wehrle/wehrle.htm

[2] Geneva Emotion Research Group:
http://www.unige.ch/fapse/emotion/members

[3] The *Emotion Home Page* maintained by Jean-Marc Fellous and Eva Hudlicka
at the Salk Institute, Computational Neurobiology Lab:
http://emotion.salk.edu/emotion.html

[4] Affective Computing MIT Media Laboratory:
http:www-white.media.mit.edu/vismod/demos/affect/affect.html

[5] The Berkeley Psychophysiology Laboratory at the University of California:
http://socrates.berkeley.edu/~ucbp1/bpl.html