Designing Emotions for Activity Selection

Dolores Cañamero*
LEGO Lab, Department of Computer Science
University of Aarhus
Aabogade 34
DK–8200 Århus N, Denmark
lola@daimi.au.dk

What emotions are about is action (or motivation for action) and action control.

[Frijda 1995, p. 506]

Abstract

This paper advocates a "bottom-up" philosophy for the design of emotional systems for autonomous agents that is guided by functional concerns, and considers the particular case of designing emotions as mechanisms for action selection. The concrete realization of these ideas implies that the design process must start with an analysis of the requirements that the features of the environment, the characteristics of the action-selection task, and the agent architecture impose on the emotional system. This is particularly important if we see emotions as mechanisms that aim at modifying or maintaining the relation of the agent with its (external and internal) environment (rather than modifying the environment itself) in order to preserve the agent's goals. Emotions can then be selected and designed according to the roles they play with respect to this relation.

1 Introduction

Autonomous agents are embodied, natural or artificial systems in permanent interaction with dynamic, unpredictable, resource-limited, and in general social environments, where they must satisfy a set of possibly conflicting goals in order to survive. Emotions—at least a subset of them—are one of the mechanisms found in biological agents to better deal with such environments, enhancing their autonomy and adaptation. Following a biomimetic approach that takes inspiration from mechanisms existing in natural systems can be a very useful principle to design emotion-like mechanisms for artificial agents which are confronted with the same kind of environments and tasks.

*On leave from the Artificial Intelligence Research Institute (IIIA) of the Spanish Scientific Research Council (CSIC). E-mail: lola@iiia.csic.es. URL: http://www.iiia.csic.es/~lola

The notions of autonomy and adaptation can be understood in a number of different ways, but we will restrict ourselves to considering the sense they have within action selection problems and how emotions relate to them in this particular context. Agents can be autonomous at different levels, depending, among other things, on their morphology, on their functional and cognitive complexity, on the features of the environment in which they are embedded, and on the kind of interactions they entertain within it. Different "levels" of autonomy can be paired with different "levels" of behavioral complexity, and we suspect that emotions might not be present nor relevant at every level. To begin with, we say an agent is autonomous if it can act by itself, without needing to be commanded by another agent. This type of autonomy can be put in correspondence with very simple behavior, such as reflex behavior responding to external stimuli alone, as in the case of purely reactive, environment-driven agents. No role seems to be left or needed for emotions in this very simple stimulus-response schema. A higher degree of autonomy is achieved when the agent can "choose" whether or not to pay attention and react to a given environmental stimulus, as a function of its relevance for the agent's behavior and goals. This implies that the agent has, or can be attributed, some internal goals or motivations that drive its behavior—it is thus *motivationally autonomous*. This is the notion of autonomy we will be concerned with in this paper and in relation to which emotions start to make sense.

One of the main problems that motivationally autonomous agents are confronted with is *activity—action, behavior—selection:* how to choose the appropriate behavior at each point in time so as to work towards the satisfaction of the current goal (its most urgent need), paying attention at the same time to the demands and opportunities coming from the environment, and without neglecting, in the long term, the satisfaction of the other active needs, so that survival is guaranteed. Taking inspiration from ethology, the "classical" solutions proposed within (embodied) Artificial Intelligence (AI) to design action selection mechanisms for artificial autonomous agents or *animats* rely on reactive behaviors, responsive to external stimuli (e.g., [Brooks 1986]), or on a repertoire of consummatory and appetitive behaviors whose execution is guided by internal needs or motivations (e.g., [Maes 1991, Donnart & Meyer 1994, Steels 1994]). My claim is that, although these models can work well for moderately dynamic environments, they are less appropriate when dynamism increases. In particular, they are not very well suited to deal with contingencies, coming from the external or the internal environment, requiring an urgent response. These situations are better dealt with when a further level of autonomy exists that allows the agent to freed itself from the demands of internal needs in order to interrupt ongoing behavior and change goal priorities when the situation requires it—therefore, to adapt quickly to rapidly changing circumstances. Some simple emotions play this role in biological agents, as we will see in Section 2.1.

In the last years, some agent architectures have been proposed that integrate a set of survival-related "basic" emotions[1] as an additional element for behavior selection (e.g., [Cañamero 1997a, Velásquez 1998]). This paper reflects on some of the lessons learned from this endeavor to point out a number of issues and problems that need to be taken

[1]The use of the expression "basic emotions" is highly controversial in the psychological literature, as authors do not agree neither on the subset of emotions that can be considered as basic, nor in what sense they are such. See [Ortony *et al.* 1988] for a very good presentation of this controversy.

into account when designing emotions for action selection.

2 Which Emotions?

The term "emotions" covers a wide range of complex phenomena which are difficult to group under a common definition. Emotions have many facets and can be approached from different levels of explanation, as they encompass neuro-endocrine, physiological (visceral and muscular), behavioral, cognitive, communicative, and social aspects. All these aspects are not always active in every emotion episode, nor in every emotion. Indeed, some emotions seem to involve strong physiological and behavioral manifestations, with weak or largely absent cognitive aspects; on the contrary, the cognitive element seems to be predominant in other emotions, which would then be close to thoughts, and that might (apparently) lack any physiological correlates; still other emotions seem to be predominantly social; also, biological and cultural influences differ considerably among different emotions. It thus seems unlikely that all this structural diversity (not to mention the functional one) can be easily accounted for by a single definition. A metaphor could perhaps be more helpful at this point. Emotions may conveniently be thought of as complex dynamic processes that integrate various causally related subsystems, as put forward for instance by [Mandler 1985, Pfeifer 1991], namely the physiological, the cognitive-evaluative, the communicative-expressive, and the subjective experience subsystems. Other important subsystems equally involved in emotion seem to have been left out from this list, such as the motivational one, but the general idea of a dynamic system involving processes of causally related subsystems seems very promising in accounting for the diversity and complexity of emotions.

In this paper, I will not deal with emotions at this general level, though; nor will I consider the different kinds of emotions mentioned in the previous paragraph, but will restrict myself to analyzing the features of some survival-related simple emotions which are involved in activity selection, and the advantages and drawbacks of taking inspiration from these features to design action selection mechanisms for artificial autonomous agents.

2.1 Natural emotions and action selection

Activity selection is a kind of adaptation that occurs over short periods of time in highly dynamic environments, to timely face rapid and usually temporary changes in the environment. It might be accompanied and improved by learning—a type of adaptation that occurs over longer periods of time—but in principle learning is not necessary for action selection to happen. Action selection in the problem faced by a motivationally autonomous agent that must choose the appropriate behavior at each point in time so as to work towards the satisfaction of the current goal (its most urgent need), paying attention at the same time to the demands and opportunities coming from the environment, and without neglecting, in the long term, the satisfaction of the other active needs, so that survival is guaranteed. Motivational and emotional states can be considered to play complementary roles in this activity.

Motivational states, such as hunger, thirst, aggression, social attachment, etc., are

drives that constitute urges to action based on internal needs related with survival. Motivations involve arousal and satiation by a specific type of stimulus, and vary as a function of deprivation. They have three main functions [Kandel *et al.* 1995]: (1) they steer behavior toward, or away from, a specific goal; (2) they increase general alertness and energize the individual to action; and (3) they combine individual behavioral components into a goal-oriented sequence. They are thus stimulus-specific, requiring urgent satisfaction, and are goal- or task-oriented. In action selection, motivational states guide the choice of the behavior(s) that must be executed—those best contributing to the satisfaction of the most urgent current need(s). They can be thought of as being concerned with appetitive processes that try to activate action as a response to deprivation; emotions, on the contrary, would be rather related with processes that try to stop ongoing behavior and are concerned with satiety processes of re-equilibration [Pribram 1984]. We could say that while motivations are in charge of driving behavior under normal circumstances, emotions take over behavior control and change goal priority in situations requiring an urgent response. But even though motivations and emotions may play complementary roles with respect to action selection, they cannot be placed at the same level. According to Tomkins, emotions, unlike motivations, show some generality of object and time, which makes that a person can experience the same emotion under different circumstances and with different objects, and, in complex systems, they amplify the effects of drives, which have insufficient strength as motives [Tomkins 1984, p. 164]: "The affect system is the primary motivational system because without its amplification, nothing else matters, and with its amplification, anything else *can* matter."

From an evolutionary perspective, emotions are biological functions of the nervous system involved in the survival of both the individual and the species in complex, dynamic, uncertain, resource-limited, social environments over which agents have little control. The emotions we are mainly considering here are the subset known in the emotion literature as "primary" or "basic", in the sense that they are more closely related with survival and are universal or very often found across cultures, which might explain the fact that they have rather distinctive physiological and expressive manifestations. Following LeDoux, each (basic) emotion evolved for different reasons and may involve different brain subsystems [LeDoux 1996]. The different emotions can be regarded as mechanisms that contribute to safeguard survival-related goals by modifying or maintaining the relation that an individual has with its (external and internal) environment in different ways [Frijda 1995]: by protecting itself from environmental influences (fear) or by blocking them (anger), by diminishing the risks of dealing with an unknown environment (anxiety), etc. Considering the emotional system as a whole, emotions have many different functions, but as far as action selection is concerned, we are mainly interested in three of them:

1. *Bodily adaptation* that allows to rapidly deal with dangers, unexpected events, and opportunities.

2. *Motivating and guiding action.* At the simplest level, the categorization of events as pleasant/unpleasant, beneficial/noxious, turns neutral stimuli into something to be pursued or avoided. As already mentioned, emotions can also change goal or motivation priority to deal with situations that need an urgent response, and amplify the effects of motivation.

3. *Signaling the relevance of events to others.* The external manifestations of emotions displayed by an individual can be taken advantage of by its conspecifics to recognize its emotional state and use this as social reference to anticipate its possible behavior or to assess the current situation.

2.2 Artificial emotions for action selection?

Emotions are a fact in humans and other animals. However, why would we want/need to endow artificial agents with emotions? Two main answers are possible, depending on what our principal concern is when modeling emotions. If we have a more "scientific" interest, we will use artificial agents as a testbed for theories about natural emotions in animals and humans, providing a synthetic approach that is complementary to the analytic study of natural systems. If we have a more "engineering" motivation, we might want to exploit some of the roles that emotions play in biological systems in order to develop mechanisms and tools to ground and enhance autonomy, adaptation, and social interaction in artificial and mixed-agent societies. The underlying hypothesis here is that, since in animals emotions are mechanisms enhancing adaptation in dynamic, uncertain, and social environments, with limited resources and over which the individual has a very limited control, when an artificial agent is confronted with an environment presenting similar features, it will need similar mechanisms in order to survive. In particular, as far as activity selection is concerned, we as engineers are interested in emotions as mechanisms that allow to:

- have rapid reactions (*fast adaptation*);
- contribute to resolve the choice among multiple goals (role in *motivating and guiding behavior*);
- signal relevant events to others (*expressive and communicative function*).

Deciding how to actually implement those mechanisms raises a number of issues, some of which we will examine in Sections 3 and 4.

But before proceeding with that, the very idea of a biomimetic approach to emotion modeling needs some consideration. Natural emotions are the product of a long evolution. However, the designer of an autonomous agent could develop different mechanisms, possibly more simple and optimized, in order to fulfill the same roles that emotions play in activity selection and in adaptation in general. Why then should we adopt the metaphor of emotions? My answer to this question is that emotions allow for a higher economy in design, at two levels. On the one hand, since an emotional system is a complex system connected to many other behavioral and cognitive subsystems, it can act on these other systems at the same time. On the other hand, since emotions are related with goals, rather than with particular behavioral response patterns [Frijda 1995, Rolls 1999], they contribute to the generation of richer, more varied, and flexible behavior.

Still another argument can be set forth against the use of emotion-like mechanisms for action selection in artificial autonomous agents. Indeed, natural emotions seem to be maladaptive in some occasions. Why use them at all, then, instead of designing more

optimized mechanisms that do not present this drawback? At the individual level, the one we are mostly concerned with here, emotions seem to be maladaptive especially when their intensity is too high, such as when a strong fear freezes us and prevents action, but also, let us not forget it, when their intensity is too low—e.g., the absence of emotion prevents normal social interaction, as evidenced for example by the studies conducted by Damasio and his group on patients with damage to the prefrontal cortex and the amygdala [Damasio 1994]. Other types of dysfunctionalities, such as some mental disorders, can be explained as improper synchronization of the different emotional subsystems [Rolls 1999]. It would thus seem that emotions are mostly maladaptive "at the extremes", but adaptive when the system is working under normal conditions. The designer of an artificial emotional system would therefore have two choices, depending on what use the system is intended for. We could design an emotional system that only works within the normal, adaptive range, and gets "switched off" (or "on") or inhibited when approaching the dangerous zone, although one must acknowledge that this could be rather difficult to achieve. We could also model the full working of a natural emotional system, including its dysfunctionlities, on the grounds of two main arguments. On the one hand, we could think that, even though emotions are sometimes maladaptive at the individual level, this maladaptiveness might still be adaptive in some way or another at the level of the species, and a computer simulation could help us understand how this could be. On the other hand, an artificial model comprising maladaptive emotional phenomena could perhaps shed some light on the causes and developing factors of some emotional disorders in humans.

3 Models of emotions for activity selection

First of all, I place myself within a "nouvelle"—embodied, situated—Artificial Intelligence (AI) perspective and claim that this approach to AI is better suited than "classical" or symbolic AI to model emotions as mechanisms for action selection in autonomous agents. In my view, the emphasis of situated AI on complete creatures in closed-loop interaction with their environment allows for a more natural and coherent integration of emotion (at least the "non-cognitive" or perhaps the "non-conscious" aspects of it) within the global behavior of the agent. This view has some implications concerning the way emotions are to be modeled in animats. Let us just mention two of them.

- First, the aspect of emotions that really matters here is how they affect the relationships between the agent and its environment; therefore, our model of emotions must clearly establish a link between emotion, motivation, and behavior, and how they feed back into each other.
- Second, this link must be grounded in the body of the agent—for instance, by means of a synthetic physiology [Cañamero 1997b]—since it is through the body that the agent interacts with the world.

Concerning emotions themselves, several types of models implementable from a situated AI perspective (and some of them from a symbolic AI perspective as well) and

applicable to behavior selection problems can be found in the literature. I'll classify them according to two criteria: the modeling goal and the viewpoint adopted on emotion.

3.1 The modeling goal

From this perspective, we can distinguish two "pure" models of emotions merging the two rather similar classifications of models in [Sloman 1992] and [Wehrle & Scherer 1995]: phenomenon-based/black-box models and design-based/process models. As I will argue below, both types of models are perfectly suited to be used for activity selection problems, and the choice of one or another will depend on the reason why we want to endow our agents with emotions.

3.1.1 Phenomenon-based or black-box models

These models assume (implicit or explicitly) the hypothesis that it is possible to somehow recognize the existence of an emotional state, and then measure its accompanying phenomena, such as physiological and cognitive causes and effects, behavioral responses, etc. It is these accompanying phenomena that this type of models reproduce, paying exclusively attention to the input/output relation and disregarding the mechanisms underlying this relation. They thus incorporate an explicit, "pre-wired" model of emotions and emotional components in charge of producing some outputs given certain inputs. These models respond to a purely engineering motivation. They can be very useful tools when emotions are introduced as behavior-producing mechanisms related to particular goals or tasks. Therefore, they can be successfully used for behavior selection tasks in cases when both the features of the environment and the kind of interactions the agent can have with it are not too complex, rather well known, and determined in advance by the engineer of the system. In this case, however, the adaptive character of emotions and their roles are taken for granted from the outset; therefore, the model cannot shed any light on the reasons and origins of the adaptive roles of emotions.

3.1.2 Design-based or process models

They pay attention to the way a system must be designed in order to produce a certain behavior, i.e., the underlying mechanisms that allow for that behavior to emerge. These mechanisms can be biologically inspired or not, and follow a top-down (e.g., [Velásquez 1996]), a bottom-up (e.g., [Pfeifer 1993]), or a middle-out approach (e.g., [Cañamero 1997a]). What really matters from this perspective is to establish a relation between the underlying mechanisms, the resulting behavior, and the environment where this behavior is situated, so as to better assess the applicability of the model. Within this perspective, it is thus possible to elaborate different emotion-based behavior selection models using different mechanisms in order to assess their particular contributions and applicability, and to compare them as a first step towards progressively achieving a higher level of generalization in the understanding their roles. Design-based models respond thus to a more "scientific" preoccupation—either because we use our artificial setting as a testbed for theories of emotions in natural systems, or because, even in the case when our main concern is to solve an AI or robotics problem, we hope to provide

some feedback regarding the equivalent problem in biological systems. As for their use to design action selection mechanisms, it can be more difficult to come up with the "appropriate" or expected behavior in each particular case, and it is not likely that these models are as robust as black-box ones, given their more exploratory nature. However, they are more useful in cases when neither the features of the environment nor the interactions of the agent with it are defined in advance or completely known, and in particular when one of our goals is to understand the relationship of emotions to both.

3.2 The view on emotion

From this perspective, we can distinguish *structural* models that split emotions into a set of components, and *functional* models, more interested in the adaptive, survival-related roles that emotions play in the interactions of the agent with its environment. Contrary to the previous classification criterion, I don't consider that these two types of models are equally adequate to guide the design of an action selection mechanism.

3.2.1 Component-based models

These models postulate that an artificial agent can be said to have emotions when it has a certain number of components characterizing human (or animal) emotional systems. Picard proposes one of such models with five components which are present in in healthy human emotional systems, although not all the components need to be active at all times when the system is operating [Picard 1997]:

1. Behavior that an observer would believe to arise from emotions (emergent emotions, emotional behaviors, and other external expressions).

2. Fast, "primary" emotional responses to certain inputs.

3. Ability to cognitively generate emotions by reasoning about situations, standards, preferences, or expectations that concern its goals in some way or another.

4. Emotional experience, i.e., physiological and cognitive awareness, and subjective feelings.

5. Interactions between emotions and other processes that imitate human cognitive and physical functions, such as memory, perception, attention, learning, decision-making, concerns and motivations, regulatory mechanisms, immune system, etc.

This list is not intended to be a formal model capturing the essential features of human-level emotion, but rather a sort of extensional definition of what it means for a computer to fully "have" emotions. To evaluate whether the goal of endowing a computer with emotions has been reached, Picard also proposes a separate test to assess the presence of each component. For the purposes of our analysis, two questions are of particular relevance with respect to this model:

1. Are all its components necessary for action selection?, and

2. What kind of guidance can it provide for the design of an action selection mechanism?

The full complexity of this model does not seem to be required for action selection. For this task, I would assume that component 1 is only needed in social decision-making situations, component 2 must be present, component 3 can be a plus in complex agents but in principle is not necessary, physiological awareness in component 4 is required, but neither cognitive awareness nor subjective feelings are, and component 5 is also needed for action selection. The choice of the components to be included in the emotional system depends not only on the complexity of the agent architecture and of the behavior to which it must give raise, but also on the nature and complexity of the action selection tasks that the agent will have to solve in the environment in which it is situated. A thorough analysis of action selection situations and of the types of architectures better suited for each of them is therefore needed before we can decide on a list of components relevant for a particular emotional system. As I have argued in [Cañamero 1998], another major problem with component-based models is that they leave open the problem of how the different components relate to each other and which are their relative priorities in the execution of different tasks and for survival in general. For these reasons, component-based models are underconstrained from a design viewpoint, since all the choices are left to the designer. In my opinion, it seems more appropriate to conceive the design process in the opposite direction—starting with an analysis of the requirements that the environment, the action-selection task, and the agent architecture impose on the emotional system. In other words, the choice of "emotional components" must be guided by functional criteria.

3.2.2 Functional models

Functional models pay attention to the properties of the structure of humans (and more generally animals) and their environment that can be transposed to a structurally different system (system = agent + environment) in order to give rise to the same functions or roles. One example of such properties that is most appropriate for action selection is provided by Frijda [Frijda 1995]:

- First, humans can be said to have the following properties relevant for the understanding of emotion:
  - They are autonomous;
  - they have multiple ultimate goals or concerns;
  - they possess the capacity to emit and respond to (social) signals; and
  - they possess limited energy- and information-processing resources and a single set of effectors.
- The human environment presents the following characteristics relevant to emotional systems:

  - It contains limited resources for the satisfaction of concerns;
  - it contains signals that can be used to recognize that opportunities for satisfaction or occurrences of threats might be present;
  - it is uncertain (probabilistic); and
  - it is in part social.

- From these characteristics (relevant to the understanding of emotion) of the human system and environment, Frijda concludes that the functions of emotion are as follows:

  - To signal the relevance of events for the concerns of the system;
  - to detect difficulties in solving problems posed by these events in terms of assuring the satisfaction of concerns;
  - to provide goals for plans for solving these problems in case of difficulties, resetting goal priorities and reallocating resources; and
  - to do this in parallel for all concerns, simultaneously working toward whatever goal the system is actively pursuing at a given moment and in the absence of any actually pursued goal.

These functional requirements are much more specific than the "components" ones, and therefore seem to be easier to attain, although they still need to be refined by adding the elements of the agent-environment system which are specific to the particular action selection situations that the agent will have to solve (e.g., which are the particular goals of the agent, what kind of signals can be exploited from the environment, etc.). Functional models also leave more freedom concerning the particular elements to be used in order to achieve these functionalities. However, as Frijda himself points out, this model remains underspecified as far as the underlying architecture and implementation mechanisms are concerned, and many problems and design issues remain open.

4 Design Issues

In his "New Fungus Eater" or emergent view of emotions, Pfeifer claims that all controversies and open problems surrounding the characterization of emotions (such as the identification of the components of emotions, the debate on the existence of a set of basic emotions, or the issues of their origins, activation, and roles) are due to the adoption of an inadequate modeling approach for their study. According to him, "these controversies *need not to be resolved*: if the approach is appropriate, they largely 'disappear' or are resolved automatically." [Pfeifer 1993, p. 918]. Most existing models up to that date being phenomenon-based, he proposed the adoption of a bottom-up, design-based modeling approach, where emotional behavior is an emergent phenomenon in the eye of the beholder.

Although I agree that design-based modeling and its principle of going below the level of the phenomena we want to study is the most appropriate for the understanding of

the origins and adaptive value emotions in action selection, Pfeifer's position presents two major drawbacks. On the one hand, modeling and implementing emotions as purely emergent phenomena in the eye of the beholder rather than as an integral part of the agent architecture can be an exciting challenge for the designer of a robot, but this view misses what I consider two of the most important contributions of artificial emotional systems: their ability to relate and influence different behavioral and cognitive subsystems at the same time, and the feedback they can provide to understand the mechanisms involved in natural emotions. On the other hand, the claim that this approach dissolves all the problems seems really extreme. At best, it can dissolve (some of) the problems posed at the "phenomenon level", but it moves the problems "down" to the design level. In particular, a fundamental problem arises from the outset: the elaboration of an "emotional agent" must be guided by a design concern, but what are the criteria that will guide our design choices? As a general guideline, I propose the two following principles:

1. Design must be guided by the following question: What does this particular (type of) agent need emotions for in the particular (type of) environment it inhabits?

2. One should not put more emotion in the agent than what is required by the complexity of the system-environment interaction.

These ideas are however too general and leave open many design problems. In [Cañamero 1998], I examined some of the questions that the functional view of emotions leaves unanswered with respect to the design of artificial emotional systems in general, namely concerning the level and complexity of the model, the controversy between engineering versus evolving the emotional system, and the evaluation of models of emotions and the contributions of these to the agent performance. Let us now consider some of the problems that must be taken into account when designing emotions for activity selection.

4.1 Study of environments

A thorough comprehension of the role of emotions in activity selection requires to understand the precise relationship between the emotional systems and the problems each contributes to solve. These problems are best characterized in terms of types of environments that allow to understand how and why the different emotions emerged and are adaptive in a particular context. A classification of environments in terms of features which are relevant for action selection—dynamism, uncertainty, threats, availability of resources, social interactions, etc.—is thus a necessary first step towards a principled design of emotional systems contributing to this task. These features can provide good guidelines to design emotional systems for action selection at least in three main respects.

In the first, and perhaps most trivial one, they provide good clues to decide what particular emotions the agent needs in that particular context. This is the case because environmental features define the kinds of problems that the agent is confronted with, and the activities or competences it must incorporate to deal with them. In doing so, they also indirectly indicate what functions or mechanisms are needed to ensure survival in that environment: protection from (particular types of) dangers, removal of obstacles,

attachment to conspecifics, etc. In this case, therefore, emotions are selected on the grounds of their survival-related functions.

Second, the characteristics of the environment allow us to assess how important each emotion is in that context, and therefore how "easily" or how often they should be activated (i.e. what the activation threshold should be for each emotion). If we follow Frijda to see emotions as mechanisms that serve to safeguard own "concerns"—goals, motivations—by modifying or maintaining one's relation to the environment [Frijda 1995], rather than as mechanisms aiming at modifying the environment itself, then the significance of the diverse types of relations one can have with the environment will depend to a big extent on the features present in it. Some of the modifications mentioned by Frijda include protection from influences from the environment in the case of fear, blocking these influences in that of anger, diminishing the risks of dealing with an unknown environment in that of anxiety, etc. The weight that each of these mechanisms has in a particular type of environment can be either established by the designer of left for the agent to learn as a consequence of its actions.

Third, the characteristics of the environment help us decide the "cognitive complexity" required for the emotional system, i.e. which "components" can be meaningfully included. For example, in a moderately dynamic environment where resources have relatively stable locations, the use of an "emotional memory" of places where different objects are likely to be found (e.g., where to seek shelter when faced with a fearful situation) can be much more efficient than a blind search in spite of its additional cognitive/computational cost. On the contrary, if the environment is so highly dynamic that objects only remain at a particular location for a short period of time, this type of memory can be more inefficient than random search. It can even have a negative effect with respect to survival, since the time used to recall and reach the position of an object that is likely to have moved could have been better used to execute a different behavior.

4.2 Choice of primitives

Emotions are included with a purpose in mind, and therefore we must carefully choose the mechanisms we need for the system to display the desired behavior, and at what level it is meaningful to model them within our architecture. When emotions must play a role in behavior selection, are these mechanisms better thought of in terms of continuous dimensions or discrete categories, to map the usual distinction in emotion theory? The answer to this question largely depends on the type of model we are using: black-box/phenomenon-based, where we must explicitly include in the model the elements that will produce the desired behaviors, or process/design-based, where the model must go below the level of the phenomenon under study to allow emotional behavior naturally emerge out of the system's interactions with its environment. For the first type of models, a "predefined" set of "discrete" emotions seems the most natural choice, but this poses the problem of which ones to choose and how "realistic" our model of them must be. As we have seen in Section 4.1, the characteristics of the environment can help us decide which emotions are relevant in each particular case. As for how "realistic" our model should be, the answer largely depends on what our agent architecture is like, and what is our purpose in building it. For the second type of models, emotional behavior could perhaps be better grounded

in some underlying simple mechanisms or "internal value systems" [Pfeifer 1993] encoding features along some dimensions. Behavior that could be thought of as arising from emotions would then be an emergent property of the agent-environment interaction. One could also think of encoding those features in a genotype that would lead to "emotional behavior" (the phenotype) and that could be evolved to generate agents with emotion-based action selection systems adapted to different types of environments. However, the problem of selecting the right primitives (the genes in this case) to give rise to meaningful behavior is still present; it could even be harder in this case, as the genotype-phenotype relation is not a linear one, and therefore not fully understandable.

In my opinion, however, presenting the discrete categories/continuous dimensions characterizations of emotions as conflicting ones is an ill-posed problem that opposes classifications belonging to different realms. Dimensions such as valence and arousal are "structural" features which are present in all emotions. If we implement emotions in our system as discrete categories, they must also possess these properties in order for them to function as natural emotions do. I would thus say that they are necessary features of emotions; however, they are not sufficient to characterize emotions since they leave out function—what is each emotion for in this agent-environment system? what are their precise roles in action selection? And function is precisely the main criterion underlying classifications of emotions in terms of categories. The two characterizations are thus complementary, and both must be taken into account, explicitly or implicitly, when designing the action selection mechanism, regardless of the primitives we choose to implement emotions.

4.3 Connection with motivation and behavior

How to relate emotions with the other two main elements involved in activity selection, namely motivation and behavior? Let us consider behavior first. The connection between emotions and behavior is not a direct one. As pointed out by several authors (see for instance [Frijda 1995, Ortony *et al.* 1988, Rolls 1999]) the fact that emotions are related with goals, rather than with particular behavioral response patterns explains the fact that they contribute to the generation of richer, more varied, and flexible behavior. Some emotions, and this is the case in relation with action selection, show characteristic action tendencies, to put it in Frijda's terms, but this relation is far from being automatic or reflex-like. The relation between emotions and behavior is thus through goals or, in the context of action selection, motivations.

Some authors (e.g., [Pfeifer 1993]) place (basic) emotions at the same level as motivations or drives, but I consider emotions as second-order modifiers or amplifiers of motivation, following for example [Tomkins 1984, Frijda 1995]. In an action selection system, motivations set goal priorities and guide the choice of the behavior(s) to be executed—those best contributing to the satisfaction of the most urgent current need(s). What role is then left for emotions in action selection? Quoting Frijda, "emotions alert us to unexpected threats, interruptions, and opportunities." [Frijda 1995, p. 506]. For this, they may interrupt ongoing behavior by acting on (amplifying/modifying) motivation, as in some cases this can imply resetting goal priorities. Emotions, at least as far as action selection is concerned, are better seen as a "second-order" behavior control or monitoring mechanism acting on top of motivational control to safeguard survival-related

concerns by modifying or maintaining our relation to the environment. As argued in [Cañamero 1997a, Frijda 1995], in order to preserve the generality feature of emotions (its task-independence) the connection between motivations and emotions must be indirect, with emotions running independent of and in parallel with motivations. One could in principle imagine different mechanisms for this indirect connection. The solution I have proposed, as we will see in the next section, relies on the use of a synthetic physiology through which these elements interact.

5 Behavior Selection using Homeostasis and Hormones

The architecture I proposed in [Cañamero 1997a, Cañamero 1997b] relies on both motivations and emotions to achieve behavior selection. The simulated creatures, with very limited perception and action capabilities, have a synthetic physiology (variables controlled homeostatically and hormones) that constitutes their internal state and makes them have motivations controlled in a homeostatic way. Controlled variables activate motivations—aggression, cold, curiosity, fatigue, hunger, self-protection, thirst, and warm—driven by internal needs or drives when their value goes out of the viability range. The motivation with the highest intensity will try to execute behaviors that can best contribute to satisfy the most urgent need(s), consummatory—attack, drink, eat, play, rest, walk, and withdraw—if the stimulus is present, appetitive—look-for, look-around, avoid, etc.—otherwise. The execution of a consummatory behavior affects both the external world and the creature's physiology. The creatures can also enter in different emotional states as a result of their interactions with the world, affecting their physiology, attention, perception, motivation, and behavior.

I will briefly examine here the design choices underlying the emotional system in terms of the issues presented in the previous section.

5.1 Design choices

5.1.1 Features of the environment

The microworld, Gridland, is a two-dimensional toroidal grid containing resources (food and water), obstacles, and two species of creatures—Abbotts (the ones integrating emotions) and Enemies (predators with a very simple behavior and fixed arbitration mechanism). It presents the following features relevant for action selection.

Dynamism. Gridland is a highly dynamic environment in which all objects can change locations. The only memory that Abbotts have is thus of types of objects encountered in the past, but not of their location.

Availability of resources. A number of food and water sources determined by the user is placed (randomly or at selected locations) all over the microworld at the beginning of a run. Food and water can be consumed by both species, and several creatures can eat or drink from a source at the same time. When a source is exhausted, it disappears and a new one is created at a random location. Resources are thus never absent from the

world, but their amount and location changes continuously, and they might be missing in a particular area when needed.

Uncertainty. It arises from two main sources: limited and noisy perception (uncertainty about the identity of objects), and the high dynamism of the world (uncertainty about the presence and location of objects beyond the small area currently perceived).

Threats. These can be present either in the external environment (Enemies, angry Abbotts) or in the internal one (physiological parameters reaching values that menace survival).

Social interactions. Most interactions, both intra- and inter-species, are either competitive (consuming the same resource) or aggressive (flee or fight). Abbotts can get happy and play in the presence of a conspecific, but this does not currently lead to cooperation nor to attachment behaviors.

5.1.2 Choice of primitives

The model includes explicit mechanisms for emotions. These are characterized by: a triggering event; an intensity proportional to the level of activation; an activation threshold; a list of hormones which are released when the emotion is active; and a list of physiological manifestations. A subset of discrete categories corresponding to "primary" or "basic" emotions were chosen that function as monitoring mechanisms to deal with important survival-related situations that arise in the relations the creatures entertain with their environment, namely

- *Anger*: A mechanism to block the influences from the environment by suddenly stopping the current situation. Its triggering event is the fact that the accomplishment of a goal (a motivation) is menaced or undone.
- *Boredom*: A mechanism to stop repetitive behavior that does not contribute to satisfy the creature's needs. Its triggering event is prolonged inefficient repetitive activity.
- *Fear*: A defense mechanism against external threats. Its triggering event is the presence of Enemies.
- *Happiness*: It is a double mechanisms. On the one hand, a re-equilibration one triggered by the achievement of a goal. On the other hand, an attachment mechanism triggered by the presence of a conspecific, although this second mechanism is not further exploited in the current implementation.
- *Interest*: A mechanism for the creature to engage in interaction with the world. Its triggering event is the presence of a novel object.

- *Sadness*: A mechanism to stop an active relation with the environment when the creature is not in a condition to get a need satisfied. By slowing down the creature (its metabolism and motor system) it prevents action for a while, waiting for the world or the internal state of the creature to change. Its triggering event is the inability to achieve a goal.

These discrete categories, however, have also properties of valence (through the release of hormones that act as "pain" or "pleasure" mechanisms) and arousal (physiological activity).

5.1.3 Connection with motivation and behavior

Emotions are activated either by significant events or by general stimulation patterns (sudden stimulation increase, sustained high stimulation level, sustained low stimulation level, and sudden stimulation decrease), and are discriminated by particular patterns of physiological parameters specific to each emotion. The model allows to activate several emotions at the same time [Cañamero 1997b], all of which influence behavior to various degrees through hormone release, or to adopt a winner-takes-all strategy [Cañamero 1997a] where a single emotion defines the affective state of the creature. Specific hormones that selectively modify the levels of controlled variables (and, to a bigger extent, the readings of the sensors tracking these variables, to reflect the fact that visceral effects of emotions are usually much slower than behavioral ones) are released by the different emotions when they are active. In addition, other hormones are emotion-independent, and they are released as a function of arousal.

Emotions run in parallel with the motivational control system, continuously "monitoring" the external and internal environment for significant events. Connection with behavior is through their influence on motivation. Emotions modify motivations as follows. The effects that hormone release has on the values of the controlled variables is computed before motivations are assessed, producing a modification of the creature's motivational state. The error signals that motivations have can then be different from those they would have had in the absence of emotions, and therefore their activation level or intensity will be different as well. This can either change the priority of motivations (control precedence), which in turn modifies what the creature attends to and what behavior gets selected—i.e. ongoing behavior is interrupted and a new one is selected to rapidly deal with a (external or internal) thread, an inadaptive condition in behavior execution, or an opportunity; hormonal modification of controlled variables can also change the way in which a behavior is executed—its duration or its motor strength, depending on the behavior—in the case of a modification of the motivation's intensity, as this intensity is passed to behaviors.

In addition to modifying the motivational state, emotions also influence the creatures' perception of both the external world and their own body. The former is achieved by hormonal modification of the "vigilance threshold" parameter in the ART–1 neural network responsible for forming and remembering object categories; this leads, e.g., to a coarser granularity in the categories formed and therefore to a "confused" state when emotions are active with a very high intensity, or to finer categorization under an "alert" state (moderately high emotion activation). Altered body perception is achieved through

hormonal modification of the readings of internal sensors (those measuring the values of controlled variables). For example, endorphine is released under a very happy emotional state, and this reduces the perception of pain.

5.2 Looking backward, looking forward

This model meets the main requirements that action selection in a highly dynamic environment imposes on an emotional system. It includes a wide enough repertoire of mechanisms to allow the agent to modify the relation with its environment, both external and internal, in ways which are beneficial for the attainment of the agent's goals, contributing to its survival and well-being. The connection between emotions, motivations and behavior through a physiology is also appropriate for action selection, as the fact that emotion constitute a "second-order" behavior control mechanism running in parallel with motivational control allows emotions to continuously monitor and timely influence the other two elements, while keeping their

Since the relation between emotions and behavior is only indirect, through motivations, the agent can show flexible and varied behavior. As for their connection to motivations, emotions are capable of appropriately modifying (by altering motivation/goal priorities) or amplifying (by raising or lowering the intensities of motivations) the effects of motivation on behavior to ensure proper treatment of situations needing an urgent response.

One of the main problems with this architecture is that it was totally hand-coded, and therefore very difficult to tune, in particular the connections of the different elements through the physiology. Using evolutionary techniques to generate different emotional systems would therefore be a very interesting direction to explore, as it would also allow to evaluate the performance and adaptive value of emotions for different action selection tasks and environments.

The emotional system was designed to meet the requirements of action selection in a very dynamic environment, where timely and fast responses are more important than sophisticated solutions. The emotional system is thus very simple. In particular, the more "cognitive" aspects of emotions, such as anticipation, appraisal of the situation and of the possible consequences of the emotional response, control of the emotions, and emotion-related learning were not explored. These aspects can however be important for more "complex" action selection situations, in particular when social relations come into play.

6 Conclusion

In this paper I have advocated a "bottom-up" philosophy for the design of emotional systems for autonomous agents that is guided by functional concerns, and I have considered the particular case of emotions as a mechanism for action selection. The elaboration of artificial emotional systems should be guided by two main concerns:

1. Design must be guided by the following question: What does this particular (type of) agent need emotions for in the particular (type of) environment it inhabits?

2. We should be careful not to put more emotion in the system than what is required by the complexity of the agent-environment interaction.

The concrete realization of these ideas implies that the design process must start with an analysis of the requirements that the features of the environment, the characteristics of the task, and the agent architecture impose on the emotional system. This is particularly important if we see emotions as mechanisms that aim at modifying or maintaining the relation of the agent with its (external and internal) environment (rather than modifying the environment itself) in order to preserve the agent's goals of well-being. Requirements determine what kind of activities and mechanisms are needed to ensure that this relation favors the preservation of the agent's goals. Emotions can be then selected and designed according to the roles they play with respect to this relation.

In the case of action selection, the most relevant features of the environment are dynamism, uncertainty, threats, availability of resources, and social interactions. A careful analysis of these features can provide good pointers as to what kind of emotional system is needed for the agent to modify the relation with its particular environment in ways that are beneficial for the preservation of its own goals, specially in three main aspects: the choice of the emotions needed in that context, the relevance of each emotion in that precise environment, and "cognitive complexity" (emotional "components") that is meaningful for that concrete emotional system.

As for the requirements that action selection imposes on emotions, the most important consideration that needs to be taken into account is how emotions relate with motivation and behavior in order to properly deal with situations requiring a rapid response—threats, unexpected events, inefficient behavior in the pursuit of goals, etc. I see emotions as a "second-order" behavior control mechanism acting in parallel with, and on top of, motivational control, to continuously monitor the (external and internal) environment for situations in which the relation of the agent with its environment has some significance with respect to the agent's goals. They affect behavior selection indirectly, by amplifying or modifying the effects of the motivational system, e.g., resetting goal priorities and redirecting attention towards a situation that needs prompt solution. The solution I have proposed to connect these three elements—emotions, motivations, and behavior—while keeping them at different "levels" relies on a synthetic physiology through which they interact. This solution was however conceived for a particular type of action selection problem and environment, and other solutions might reveal more adequate in other contexts.

Finally, concerning the requirements that the agent architecture imposes on the emotional system, the can greatly vary depending on the type of architecture we are using, and this in turn depends on what is our purpose when endowing our agent with emotions.

Acknowledgments

I am grateful to Robert Trappl and Paolo Petta for creating the right atmosphere to meet with other researchers and exchange ideas about Emotions in Humans and Artifacts in their Lab in Vienna, and for (their patience!) encouraging me to write this paper. Discussions with the other participants at the Vienna meeting were an invaluable source of ideas.

References

[Brooks 1986] Brooks, R.A. 1986. A Robust Layered Control System for a Mobile Robot, *IEEE Journal Of Robotics and Automation*, RA–2, April: 14–23.

[Brooks 1991] Brooks, R.A. 1991. Intelligence without Representation, *Artificial Intelligence* **47**: 139–159.

[Cañamero 1997a] Cañamero, D. 1997. Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In W. Lewis Johnson, ed., *Proceedings of the First International Conference on Autonomous Agents*, 148–155. New York, NY: ACM Press.

[Cañamero 1997b] Cañamero, D. 1997b. A Hormonal Model of Emotions for Behavior Control. VUB AI-Lab Memo 97–06, Vrije Universiteit Brussel, Belgium.

[Cañamero 1998] Cañamero, D. 1998. Issues in the Design of Emotional Agents. In *Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium*. Technical Report FS-98-03. Menlo Park, CA: AAAI Press, 49–54.

[Damasio 1994] Damasio, A. 1994. *Descartes' Error. Emotion, Reason, and the Human Brain*. New York, NY: G.P. Putnam's Sons.

[Donnart & Meyer 1994] Donnart, J.Y., and Meyer, J.A. 1994. A hierarchical classifier system implementing a motivationally autonomous animat. In D. Cliff, P. Husbands, J.-A. Meyer, S.W. Wilson (Eds.) *From Animals to Animats 3. Proceedings of the Third International Conference on Simulation of Adaptive Behavior*. Cambridge, MA: The MIT Press/Bradford Books.

[Frijda 1995] Frijda, N.H. 1995. Emotions in Robots. In H.L. Roitblat and J.-A. Meyer, eds., *Comparative Approaches to Cognitive Science*, 501–516. Cambridge, MA: The MIT Press.

[Kandel *et al.* 1995] Kandel, E.R., Schwartz, J.H., Jessell, T.M. 1995. *Essentials of Neural Science and Behavior*. Norwalk, CT: Appleton & Lange.

[LeDoux 1996] LeDoux, J. 1996. *The Emotional Brain*. New York, NY: Simon & Schuster.

[Maes 1991] Maes, P. 1991. A Bottom-Up Mechanism for Behavior Selection in an Artificial Creature. In J.-A. Meyer & S.W. Wilson, eds., *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, 238–246. Cambridge, MA: The MIT Press.

[Mandler 1985] Mandler, G. 1985. *Mind and Body*. New York, NY: W.W. Norton.

[Ortony *et al.* 1988] Ortony, A., Clore, G.L., Collins, A. 1988. *The Cognitive Structure of Emotions*. New York, NY: Cambridge University Press.

[Pfeifer 1991] Pfeifer, R. 1991. A Dynamic View of Emotion with an Application to the Classification of Emotional Disorders. Vrije Universiteit Brussel, AI Memo 91–8.

[Pfeifer 1993] Pfeifer, R. 1993. Studying Emotions: Fungus Eaters. *Proceedings of the Second European Conference on Artificial Life, ECAL'93*, 916–927. ULB, Brussels, Belgium, May 24–26.

[Pfeifer 1994] Pfeifer, R. 1994. The "Fungus Eater Approach" to Emotion: A View from Artificial Intelligence, *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, Vol. 2, No. 1, 42–57 (in Japanese). Also available as University of Zurich AI Lab Technical Report 95–04.

[Picard 1997] Picard, R.W. 1997. *Affective Computing*. Cambridge, MA: The MIT Press.

[Pribram 1984] Pribram, K.H. 1984. Emotion: A Neurobehavioral Analysis. In K.R. Scherer & P. Ekman, eds., *Approaches to Emotion*, pp. 13–38. Hillsdale, NJ: Lawrence Erlbaum Associates.

[Rolls 1999] Rolls, E.T. 1999. *The Brain and Emotions*. Oxford University Press.

[Sloman 1992] Sloman, A. 1992. 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-Verlag.

[Steels 1994] Steels, L. 1994. Building Agents with Autonomous Behavior Systems. In L. Steels & R. Brooks (Eds.) *The 'artificial life' route to 'artificial intelligence'. Building situated embodied agents*. New Haven: Lawrence Erlbaum Associates.

[Tomkins 1984] Tomkins, S.S. 1984. Affect Theory. In K.R. Scherer & P. Ekman, eds., *Approaches to Emotion*, pp. 163–195. Hillsdale, NJ: Lawrence Erlbaum Associates.

[Velásquez 1996] Velásquez, J.D. 1996. *Cathexis: A Computational Model for the Generation of Emotions and their Influence in the Behavior of Autonomous Agents*. MIT Media Laboratory MSc Thesis. Cambridge, MA.

[Velásquez 1998] Velásquez, J.D. 1998. Modeling Emotion-Based Decision-Making. In *Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium*. Technical Report FS-98-03. Menlo Park, CA: AAAI Press, 164–169.

[Wehrle & Scherer 1995] 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.