Emotional Disorders in Autonomous Agents ?

Aapo Hyvärinen[1] and Timo Honkela[2]

[1] Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400, FIN-02015 HUT, Finland
aapo.hyvarinen@hut.fi
[2] Media Lab, University of Art and Design Helsinki,
Hämeentie 135 C, FIN-00560 Helsinki, Finland
timo.honkela@uiah.fi

Abstract. It has been recently suggested by a number of authors that modelling of emotions and related motivational systems in agents might have great practical value, apart from the interest of providing possible explanations for the emotional mechanisms of human agents. Emotions, or needs, may be used as signalling mechanisms between different subsystems (subagents) inside an agent, as well as between different agents. In this paper, we investigate some problems that may arise with emotional agents. Since needs and emotions are largely global, stable reaction tendencies, they may exhibit rigidities that lead to different forms of maladaptive behavior, i.e. behavior that is not well suited to the present environment of the agent. We investigate emotional learning in agents by an utterly simplified decision-theoretical model. We show that even in this very simple model agents may develop maladaptive patterns of behavior that closely resemble patterns found in emotional disorders in humans. The maladaptive behavior patterns are due to non-optimal values for the two decision parameters, which are functions of the prior beliefs of the agent.

1 Introduction

A central issue in artificial life, as well as in artificial intelligence, is the study of adaptive agents. Adaptive agents learn from their environment, possibly by constructing a model of the world, so as to maximize some desired criteria. The agents' model of the world, and the ensuing behavior tendencies are not rigid, but a product of experience and interaction with the environment. This implies that agents have different models due to differential learning by the agents. Each has its own history which has led to a different model of the agent world.

It has been recently suggested by a number of authors, e.g. [2, 6], that modelling of emotions and related motivational systems in agents might have great practical value, apart from the interest of providing possible explanations for the emotional mechanisms of human agents. For example, emotions, or needs, may be used as signalling mechanisms between different subsystems (subagents) inside an agent, as well as between different agents [5, 1, 4, 6]. A seminal proposal in this line of research was Simon's interruption theory [5], where emotions interrupt ongoing goal processing in order to direct processing resources to more urgent goals.

In this paper, we investigate problems that may arise with emotional agents. Since needs and emotions are largely global, stable reaction tendencies, they may exhibit rigidities that lead to different forms of maladaptive behavior, i.e. behavior that is not well suited to the present environment of the agent. In humans, such maladaptive emotional states, and the ensuing maladaptive behavior patterns, have been a subject of a large body of research in psychology and psychiatry. Modelling emotional learning in agents, we show how agents may develop maladaptive patterns of behavior that closely resemble patterns found in emotional disorders in humans.

We show that maladaptive emotion-based behavior can emerge from a very simple model of the agents' decision-making system. The basic emotional system in our model can be reduced to two parameters that can be learned e.g. by reinforcement learning. Non-optimal values for these parameters may lead to maladaptive behavior patterns that resemble such disorders as depression, anxiety, and mania. A cognitive interpretation of the parameters in terms of (Bayesian) prior beliefs shows how emotional disorders are related to erroneous beliefs on the world. The results are relevant to the design of emotional agents, and may give insight into the processes involved in human maladaptive behavior.

2 Minimal Agent model

To begin with, we describe an very simple agent model that we use to illustrate maladaptive emotion-based behaviors. We accomplish this with an agent that has just one long-term goal, three internal needs (emotions), two types of sensory perceptions, and three action alternatives.

Goal and needs. In our highly simplified agent model, the agent has (or seems to have) just one ultimate long-term *goal*. This is the combination of actions, states or objects that the agent needs to perform or obtain to be useful for its designer (in the case of a robot) or to spread its genes (for biological agents). Typically, the satisfaction of the goal is not a binary truth value, but rather a scalar quantity S that measures how well the agent performed during its lifetime.

Though the agent has only one (ultimate) goal, we consider that its decision-making system consists of several modules that correspond to different *needs*. (We ignore here the modules needed for perception, action etc.) The needs are closely connected to emotions because it may be considered that emotions are expressions of current needs. Thus we model emotions in our system indirectly, via the underlying needs. Biological systems show that it is useful for an agent to construct the decision-making device of components, each corresponding to one type of constraint of the environment, or one type of subgoal. A need corresponds thus in our terminology to one of those subgoals. An emotion can be considered as a signal that expresses the urgence of a given need.

Firstly, to generate behavior that is useful for directly satisfying the (long-term) goal, the agent may be considered to generate a set of internal states that we may call the *goal need*.

If the agent just had its goal need, its behavior would consist of trying to do anything that lets it satisfy its goal need at the present instant. Such a short-sighted behavior would, however, lead to rather small goal satisfactions in a hostile environment. Thus we include in our agent model a second need which consists of trying to avoid any external events that might destroy the agent. We call this need the *safety need.*

Another need that in often in conflict with the goal need is the *energy need.* This means the need to get energy when energy level is low, by means of drinking, eating, charging batteries, etc. For simplicity, we assume in our model that the agent gains energy automatically with time.

Perception of environment. When some need-relevant event occurs, we assume the perceptual system computes two conditional probabilities: First, $P(event|goal)$, the conditional probability that an event that satisfies the goal need would produce such a perception, and second, $P(event|danger)$, the conditional probability that an event that threatens the safety need, i.e. is dangerous, would produce such a perception. Such probabilities are typically produced by most pattern-recognition methods, e.g. by computing distances from a template. It is here assumed for simplicity that all events belong to two uniform categories: those satisfying the goal need and those threatening the safety need.

Actions. On the basis of the perceptual quantities, the agent must decide its course of action among the following alternatives. 1) *Approach*: This means that the event or object is explored further, and the possible increase in goal satisfaction is enjoyed. 2) *Avoidance*: This means that the agent tries to avoid the event or object. 3) *No action.* It is assumed that both actions consume the same amount of energy, whereas the no action alternative does not consume any. Since the energy supplies are constantly replenished, this means that not doing anything really means resting to get energy.

Competition of needs. Choosing among the action alternatives is here done using a simple competition of need signals. (A more sophisticated justification of this procedure is given below.) The need modules have certain weight parameters w_{goal} and w_{danger} that are based on information they have about the context and the possible urgencies of the needs. On the basis of the perceptions and these parameters, the need modules transmit their priorities or need signals in the form of quantities $w_{goal}P(event|goal)$ and $w_{danger}P(event|danger)$. In other words, they multiply the perceptual probabilities by some quantities that express the importance of the need. Similarly, the energy need transmits a signal, which does not depend on perceptions, but possibly on some internal state of the system. We can normalize the signals so that the signal given by the energy need is equal to 1 always. Thus the decision between different action alternatives is made by choosing the maximum among the quantities

$$needsignal(goal) = w_{goal}P(event|goal) \quad (1)$$

$$needsignal(danger) = w_{danger}P(event|danger) \quad (2)$$

$$needsignal(energy) = 1. \quad (3)$$

Utility-theoretical interpretation. The decision rule given above can be interpreted in the framework of utility theory. Due to lack of space, we present here only the results of the analysis, details can be found in [3]. To decide between the three alternatives given above, the agent uses simple Bayesian decision procedures. We assume that the agent has some estimates of the following quantities. First, $\Delta S(goal)$: the increase in (life-time) goal satisfaction due to encounter with a goal satisfying event. Second, $\Delta S(danger)$: the expected decrease in (life-time) goal satisfaction due to destruction caused by encounter with dangerous event. Moreover, the agent has estimated from the environment the following prior probabilities, i.e. probabilities that express the prior belief on the occurance of the two kinds of events that the agent has before observing the present environment : $P(danger)$: the prior probability that a dangerous event occurs and $P(goal)$: the prior probability that a goal satisfying event occurs.

By some algebraic manipulation, we see that the weights w_{goal} and w_{danger} are essentially given by two quantities, which are the *prior* expected utilities $P(goal)\Delta S(goal)$ and $P(danger)\Delta S(danger)$. These quantities can be interpreted as the level of danger or goal satisfaction that the agent expects to receive on the average. They contain the essence of the prior beliefs that the agents holds on the environment, and individual differences can be traced to different estimates of these prior expectations.

3 Maladaptive emotion-based behavior

By maladaptive emotion-based behavior, we mean here configurations where the emotional components of the agent do not allow as large a goal satisfaction as were otherwise possible. Such maladaptive behavior patterns are closely related to emotional disorders, sometimes called neuroses.

In our model, we have two parameters, w_{danger} and w_{goal} (and the corresponding parameters in the Bayesian framework) that govern the behavior of the agent. Suboptimal values of these parameters are the main cause of maladaptive behavior. There might be several *reasons* why the parameters have suboptimal values. First, the environment may change. In particular, the environment where the parameters were first learned may not be actual anymore. In biological agents, this phenomenon seems to be of great importance especially because many parameter values seem to be fixed during childhood learning. Second, the learning method might be faulty. This is usually not so much a problem in biological agents, where evolution has provided quite efficient learning methods, but in artificial agents, the learning algorithms are not always efficient. It is also possible that learning of biological agents might be disturbed by diseases and environmental toxins. Third, there might be discrepancies in the long-term goals. Especially in artificial agents, there may exist several possible definitions of goal satisfactions with correspondingly different optimal values for the parameters. We can identify three classical types of maladaptation, as discussed next

Anxiety. If the coefficient w_{danger} is too high with respect to both w_{goal} and the constant energy need, this leads to a preponderance of avoidance behavior,

reducing approach behavior and inactivity. Thus the agent avoids even events whose probability of being dangerous is relatively small. This parallels what is called anxiety (or the generalized anxiety syndrome) in humans.

In a cognitive interpretation, the weight w_{danger} is essentially the prior expected loss of utility due to dangerous events. An elevated value for w_{danger}, and the associated tendency to avoidance, can thus be caused by a strong belief in the occurrence of dangerous events, or on the high value placed on safety. Both of these give rise to a cognition that exaggerates the importance of dangerous events.

Depression. If w_{goal} is too low compared to the constant signal emitted by the energy need, we have a situation characterized by a relative rarity of approach behavior that is not unlike the prevalent behavioral symptom of depression. Especially when combined with a low w_{danger}, this lead ultimately to inactivity and stupor as in severe depression in humans. In the cognitive, utility-theoretic, interpretation, a very low relative value of w_{goal} may be due to a pessimistic stance in which the prior belief on goal satisfaction low, or due to the low value assigned to goal satisfaction.

Mania. The third pathological state is observed when w_{goal} is too high and w_{danger} is too low. In this state, the agent ignores dangers and is excessively active in approach behavior, which resembles the symptoms of the manic states in humans.

4 Conclusion

We modelled the motivational or emotional system of a simplified agent using a system of three needs and three behaviors. It was shown how the behavior of the agent depends basically on two parameters. Suboptimal values for these parameters lead to behavior tendencies that resemble some of the behavior patterns encountered in humans with emotional disorders. The model gives explanations for the etiology of these disorders that are not unlike those given in the psychological literature on humans, and gives a new framework for studying adaptation phenomena related to emotions.

References

1. N. H. Frijda. *The Emotions.* Cambridge University Press, 1986.
2. T. Gomi, J. Vardalas, and K.-I. Ide. Elements of artificial emotion. In *Proc. Robot and Human Communication (RO-MAN'95)*, Tokyo, Japan, 1995.
3. A. Hyvärinen and T. Honkela. Maladaptive emotion-based behaviors in autonomous agents. In *Proc. STeP'98*, Jyväskylä, Finland, 1998.
4. K. Oatley. *Best Laid Schemes: The Psychology of the Emotions.* Cambridge University Press, 1992.
5. H. A. Simon. Motivational and emotional controls of cognition. In *Models of thought*, pages 29–38. Yale University Press, 1979. Reprint, originally appeared in 1967.
6. A. Sloman and M. Croucher. Why robots will have emotions. In *Proc. 7th Int. Joint Conf. on Artificial Intelligence*, Vancouver, Canada, 1981.