An Emotional Mechanism for Intelligent Agents Inspired on a Model of Anxiety

Mauren Brenner (maurenb@hotmail.com) – RA 946310
IA718A – Introduction to Cognitive Science
FEEC – Unicamp

Abstract

This paper describes an architecture for intelligent agents that incorporates an emotional control mechanism based on the threat evaluation system proposed in a model for anxiety disorders. Evolutionary computing techniques are proposed to evaluate the effectiveness of the emotional mechanism as well as to try to find out why some individuals are more prone to anxiety than others. This can be considered as an experiment in Cognitive Science, since a model from Psychology is used as a basis for an architecture of Artificial Intelligence.

Keywords: affect, anxiety, architecture, artificial agents, attention, cognition, control, danger, emotion, evolution, fear, intelligent systems, threat.

1. Introduction

This paper describes a research project that has three objectives. First, it seeks to define an emotional mechanism inspired on a theoretical model that was devised to explain information processing biases observed in anxiety disorders. Second, it proposes to incorporate this mechanism in an agent architecture and use techniques of evolutionary computing to evaluate its effectiveness. Third, it intends to use the evolutionary experiments in an attempt to understand why some individuals are more vulnerable to anxiety than others.

Emotions stand on a critical juncture of Cognitive Science: it is the subject of study of Psychology, Neuroscience and Artificial Intelligence, and the different perspectives do not easily complement each other at the current stage. The theory of "emotional intelligence" (Goleman 1996) has become immensely popular, but it has been criticised by other researchers (Matthews et al. 2003). The neuroscientist Antonio Damasio has claimed that emotions are fundamental for reasoning (Damasio 1994), but his view relies strongly on the functions of brain areas and has been criticised by the cognitive scientist Aaron Sloman (Sloman 1998). In Artificial Intelligence, relatively "shallow" models of emotion have been used to add psychological realism to characters in computer games and other applications, most notably in the entertainment area (Sloman 2001). This contrasts sharply with the view of emotions as related to control mechanisms in intelligent systems (Sloman et al. 2003). This paper is exclusively concerned with the latter view, although it proposes the use of a theoretical model from Psychology.

One of the most important branches of research in the psychology of emotions is the study of anxiety and fear, where these are regarded as emotional states associated with the detection of potential threat or danger. These studies often seek to understand the mechanisms that cause pathological levels of anxiety, which become manifest as the so-called anxiety disorders, such as specific phobias, generalized anxiety disorder, obsessive-compulsive disorder and panic disorder (APA 1994, WHO 2006). Theoretical models have been proposed to explain how these mechanisms work and to make hypotheses as to what causes the disorders. This paper uses the model proposed by Andrew Mathews and others (Mathews & Mackintosh 1998, Borkovec 2004, Yiend 2004) as a source of requirements and properties for an emotional mechanism for intelligent systems.

The architecture for the intelligent system described in this paper follows the CogAff schema proposed by Sloman and his colleagues in the Cognition and Affect Project (Sloman 1998, Sloman et al. 2003). In particular, it is based on their particular definition of emotion as a control mechanism, reviewed in the next section.

Evolutionary computing consists of a collection of techniques inspired in Darwin's theory of evolution and natural selection. Originally intended to achieve better solutions from an initial random set by introduced mutations and then pruning the population based on fitness criteria over several generations, it has found broader uses where the common theme remains essentially the same. In this paper, evolutionary experiments shall be used to determine the relative fitness or adequacy of agents in certain environments and to try to understand how vulnerable individuals could have evolved.

2. Background

2.1. Emotions as Control Mechanisms in Intelligent Systems

Herbert Simon is usually regarded as the first to propose that emotions have a control role in intelligent systems (Simon 1967). His work was refined and extended by Sloman and others (Sloman 1998, Sloman et al. 2003). According to their definition, an emotion is a kind of "alarm" that interrupts normal processing when there is the need to attend to a special event by acting (actual behaviour) or getting prepared to act (disposition). Therefore, emotions are likely to evolve in systems that are characterized by concurrent activity of multiple components such as those described in (Minsky 1987) and (Franklin 1995), where some components perform sophisticated, time-consuming operations and are thus unable to respond to unexpected events in a timely manner unless they are

interrupted by an "alarm".

There are several different proposals for the use of emotions as control mechanisms in intelligent systems. A non-exhaustive list with twelve possibilities can be found in (Scheutz 2004). Others regard emotions as the result of adaptations to characteristics of the system and its environment, such as limitations of the system to predict contingencies in the environment, the need for management of social behaviour and interpersonal communication (Michaud et al. 2001). Some still focus on the appraisal aspect of emotions while assigning distinct "signal" and "response" roles to them (Botelho & Coelho 2001).

It is important to distinguish between the notion of emotions as control mechanisms and phenomenological views of emotions. The control role of emotions does not require that these be accompanied by phenomenological experience at all. It is thus possible for an artificial intelligent system to have emotional control mechanisms without ever having any "experience" of emotions, affect or feelings whatsoever. The agents described in this paper indeed do not have such experiences.

2.2. The Architectural Approach

Due to the difficulties to reach appropriate definitions of emotions, Sloman suggests that one should adopt a design-based or architectural approach and design and implement computational models that capture essential properties of emotions. He claims that the architectural approach helps clarify pre-theoretical concepts and make them more precise to express scientific theories and engineering objectives by exposing the underlying mechanisms, processes and states (Sloman et al. 2003).

In order to support such design-based endeavours, Sloman and his group created the CogAff schema for intelligent system architectures. The CogAff schema consists of superimposed layers of processing that help characterize the components of the architecture and the interactions between them. First, components are divided into the three layers that correspond to *perception*, *control* and *action*. Second, components can belong to a *reactive* layer, to a *deliberative* layer or to a *meta-management* layer, where reactive components provide immediate responses, deliberative components execute plans and more sophisticated information processing, and meta-management components handle the allocation of resources in the system. A summary of the CogAff schema is presented in (Sloman 1998) and a longer discussion can be found in (Sloman 1999). In this schema, emotions usually have reactive components that trigger "alarms (Sloman 1998).

2.3. Information Processing Biases in Psychopathology

Throughout the last few decades it has become established that anxiety disorders are characterized by *attentional* and *interpretive biases* in information processing whereby anxious individuals are more likely to be distracted by threat cues and have a tendency to interpret ambiguous information in a more negative way (Yiend 2004). These biases in information processing have been empirically verified by means of techniques such as the Stroop task and variations, the attentional probe task (also known as the dot-probe task) and word recognition tasks using homophones (*brews/bruise*) and homographs (*stroke* can mean *brain hemorrhage* or *caress*). Brief descriptions of these techniques can be found in (Fox 2004), (MacLeod et al. 2004) and (Richards 2004).

2.4. Cognitive Models of Processing Biases

Several models of cognitive processing have been proposed to explain the processes underlying attentional and interpretive biases and account for the differences found between anxious and non-anxious subjects. Some of these models are surveyed in (Mathews & Mackintosh 1998), where a new model is proposed to address the shortcomings of the earlier ones. The model proposed by Mathews is also summarized in (Borkovec 2004).

Mathews' model describes the mechanisms of danger and threat detection involved in the emotional states of anxiety and fear in order to provide a theoretical explanation for attentional and interpretive biases observed in experiments. The model summarized here is the one described in (Mathews & Mackintosh 1998).

The model assumes that there are representations in the mind that correspond to qualities of objects (external or internal). So, for instance, in the Stroop task there are two separate representations associated with a word written in a certain colour: the colour and the meaning of the word.

Attention is a general name for mechanisms that give priority to certain representations over others, according to their level of activation. A representation can be activated because it is the focus of a task that the subject is performing. Conscious, voluntary effort can activate such representations to the level required by the task, regardless of emotional content. A representation can also be activated because it is associated with potential danger. A *threat evaluation system* (or TES) is responsible for recognition and activation of such representations. In this case the representations are said to have emotional content.

The threat evaluation system (TES) is the central mechanism in fear and anxiety response. Due to the importance of these responses for survival, the TES must operate in parallel with other cognitive functions and must be capable of rapid detection of threat cues. The TES is therefore a non-conscious, automatic mechanism. It uses a "quick-and-dirty" pattern matching mechanism that compares representations to stored patterns associated with danger. Some patterns are innate and have evolved due to evolutionary pressure, such as those associated with predators (in all animals) and social threat (in humans); other patterns can be acquired by conditioning and learning.

The effect of the TES is to activate those representations that are associated with threat. This effect depends on the *sensitiveness* of the TES: the more sensitive it is, the more likely it is that representations even remotely matching the stored patterns will be activated. The sensitiveness of the TES is clearly also related to the number of patterns available to the pattern matching mechanism. Overall TES sensitivity is correlated to the so-called *trait anxiety*, a relatively stable personality trait that does not depend on the current situation. This explains why high trait-anxious individuals are more likely to be distracted by threatening cues.

The effect of the TES is also modulated by the current fear or anxiety level, or *state anxiety* (as opposed to trait anxiety). State anxiety produces autonomic responses and increases the sensitivity of the TES. Moreover, it increases the level of activation that the TES will assign to representations of potential danger. This is why low trait-anxious individuals will respond to highly threatening information, and why even high trait-anxious people will not respond to insignificant dangers if they are not in an anxious state.

Representations that are activated by the TES can either produce an autonomic response (thus increasing state anxiety), such as startle, or compete for attention with other representations.

Representations that require common processing resources inhibit each other. This allows for a top-down control process, such as conscious effort to concentrate on a task, to direct attention to its target representations and oppose the distracting effect of a relatively weak threat stimulus. In addition to this, stronger threat representations inhibit weaker ones, so that an insignificant threat fails to capture attention when the individual is facing a more severe danger.

The TES is also called "behavioural inhibition system" because it interrupts current behaviours, initiating physiological arousal and directing attention to the potential source of danger (Mathews & Mackintosh 1998).

The pattern matching mechanism used by the TES is relatively imprecise, because the consequences of not reacting quickly enough to danger can be much worse than the consequences of overreacting to something unimportant. LeDoux suggests that stimuli can be simultaneously evaluated by the quicker, nonconscious subcortical pathway that involves the thalamus and the amygdala (and, cognitively speaking, emotional memory and the TES) and a slower, cortical route which also involves the hippocampus (and thus declarative memory) and that can stop an inappropriate response initiated by the quicker route (LeDoux 1994).

2.5. What are Emotions Useful For?

Emotions as control mechanisms have clear evolutionary advantages not only in human beings, but in other natural organisms as well (Pinker 1997, LeDoux 2002, Minsky 2006). As for artificial intelligent systems, evolutionary computing techniques have been employed by Scheutz and his colleagues to compare the performances of affective, reactive and deliberative agents in different environments (Scheutz & Logan 2001, Scheutz & Sloman 2001) and to compare emotional vs. non-emotional and social vs. asocial agents (Scheutz 2004).

3. Designing and Evaluating Emotional Agents With a Threat Evaluation System

3.1. Specifying an Architecture Based on Mathews' Model

This section describes how the essential components of Mathews' model shall be mapped into architectural properties. This description uses the terminology from the CogAff schema (Sloman et al. 2003).

The architecture proposed is as simple as possible and does not intend to be a model of the human mind. Rather, it focuses on the threat evaluation system, and shall be used in evolutionary experiments to evaluate the advantages of such a mechanism and the emergence of more "anxious" agents. The evolutionary experiments are described in the next subsection.

The threat evaluation system (TES) shall be a component in the reactive layer. It contains the pattern matching mechanism and the emotional memory that contains the patterns. This architecture shall not include any mechanism for learning new patterns; the only available ones shall be predefined.

In order to provide a counterpart for the emotional mechanism that activates distracting threat representations, the architecture shall have a component in the deliberative layer that looks for rewards in the environment and makes plans to reach them. The steps in the plan shall provide the target representations that occupy the focus of attention unless a distracting representation captures it and interrrupts the execution of the plan.

Competing representations shall be kept in an ordered list where the most active representation always has the focus of attention. It is worth noting that if there is only one representation in the list, then there is no competition and the representation is processed regardless of its level of activation. This is consistent with Mathews' statement that processing biases are only seen when stimuli compete for attention (Mathews & Mackintosh 1998).

Target representations activated by the deliberative plan shall have a constant, relatively high level of activation AR. The level of activation of a threat representation AT is determined by the TES based on the following values:

- *P* is the level of significance of a potential threat stored in the emotional memory. Lethal threats shall always be able to override target representations.

- *S* is the sensitivity of the TES. It corresponds roughly to trait anxiety and shall be a variable parameter used in the evolutionary experiments to introduce variation between individuals. *S* must be greater than zero, as $S = 0$ would correspond to the agent being oblivious to danger.

- *X* is the current anxiety level. If $X = 0$, then it shall not influence the output from the TES.

AT should be computed so that it is never less than *P*. *S* has a multiplicative effect on *P*, and so does *X* if it is greater than zero. The actual formulas and values shall be defined and adjusted so as to allow the evolutionary experiments to have meaningful results.

If a threat representation is active enough, it will cause a component in the action layer to trigger a fight-or-flight response or to increase the current anxiety level.

3.2. Evolutionary Experiments

A number of agents shall be implemented in a virtual world that contains both *rewarding objects* and *aversive objects*. Rewarding objects (e.g. food or energy sources) increase the energy level of the agent that gets them, whereas aversive objects (e.g. predators, poison or traps) decrease it. Every agent shall be able to recognize rewarding and aversive objects from the outset, i.e. there will be no need to learn or to evolve it.

When the energy level of an agent drops down to zero, the agent dies. When the energy level goes up to a certain threshold, the agent can reproduce using up energy in the process. Reproducing consists in spawning a number of offspring that correspond to clones or copies with slight variations in the trait anxiety parameter *S*. The energy level of an agent at any time can thus be regarded as a measure of individual fitness.

Agents shall roam the virtual world looking for rewarding objects and following plans to reach them while trying to avoid aversive objects by means of the emotional mechanism.

Four kinds of evolutionary experiments shall be run. For the purposes of these experiments, a *catastrophe* is defined as a marked decrease in the number of rewarding objects available and/or a marked increase in the number of aversive objects. During a catastrophe period, the world becomes rather hostile for a while, which is expected to significantly reduce the agent population size. The conjecture here is that a catastrophe that decimates most of the agent population might allow highly "anxious" individuals to survive and reproduce. However, it may be the case that a catastrophe is *not* a necessary condition for this to happen. In any case, this is compatible with what is known about human evolution, i.e. that there was a period less than a hundred thousand years ago when mankind was reduced to a small population, which explains the relatively low genetic variability of our species (Pinker 2002).

The evolutionary experiments shall also employ two different kinds of agents: agents that are implemented according to the architecture defined in the previous subsection, and agents that are implemented according to the same architecture without the TES. The former are referred to as *threat-sensitive agents* and the latter as *threat-insensitive agents*.

According to the definitions above, the four kinds of evolutionary experiments that shall be run are:

1. A virtual world where threat-sensitive and threat-insensitive agents compete in the absence of catastrophes;

2. A virtual world where theat-sensitive and threat-insensitive agents compete in the presence of a catastrophe at a certain point;

3. A virtual world where there are only threat-sensitive agents and no catastrophes happen;

4. A virtual world where there are only threat-sensitive agents and a catastrophe occurs at a certain point.

Experiments with two types of agents should provide results that allow for the evaluation of the effectiveness of the emotional mechanism to survival. It is expected that the evolutionary experiments require adjustments to the architecture so the agents can survive in the virtual world.

4. Discussion

The project proposed in this paper brings up several questions regarding its validity, limitations and directions for future work. This section discusses some of the issues that are likely to arise.

First of all, there are fundamental questions regarding the choice of theoretical model and the validity of adopting a model based on pathological behaviour as an inspiration to architectural mechanisms. Why would anyone want to do that? For one thing, it is important to make a clear distinction between abnormal functioning of a system and the properties that this reveals about the system. Studies of dysfunctional behaviour in psychology lead to insight regarding the mechanisms underlying the behaviour, both in its normal and abnormal forms. Besides, more often than not it is impossible to discern a clear-cut distinction between normal and dysfunctional behaviour. Sometimes the difference is quantitative rather than qualitative (as it may well be the case here); sometimes the distinction

depends on the context, so that behaviour that is abnormal in certain situations can be adaptive or even normal in different circumstances. A special case occurs when behaviour that was advantageous for survival and reproduction at a certain point in evolution becomes maladaptive later. Finally, dysfunctional behaviour may reveal fundamental trade-offs that had to be undertaken so that the mechanism could fulfil its role; for instance, a mechanism that detects and responds to danger must be fast, but this may make it less accurate and more error-prone.

Another crucial question about the choice of model is that Mathews' model was devised to explain specific empirical findings, and it also claims to be consistent with neurophysiological theories (Mathews & Mackintosh 1998). But his model can be challenged in the future by results brought upon by novel experiments, as well as advances in neuroscientific research. On the other hand, if this ever happens, then a new theory will be eventually put forward that is not only capable of accounting for the new findings, but also the old ones (unless some fundamental flaw is found in the previous methodology). In this case, it would be interesting to investigate if and how the proposed architecture could be extended or modified to comply with the new theory without being completely reformulated. In addition to this, one could ask whether the architecture itself provided any clues to the aspects uncovered by the new scientific experiments.

A different question could be asked regarding the choice of any model that tried to explain dysfunctional behaviour in terms of information processing. In principle, dysfunctions may result from neurochemical effects in the brain that affect cognitive mechanisms, rather than from their functional properties. As an analogy, one can think of troubles of declarative memory caused by prolonged stress: in this case, stress hormones deplete hippocampal neurons (which participate in the formation of declarative memories) of glucose if they stay around for too long, thereby causing these cells to have a toxic reaction to the neurotransmitter glutamate; the neurons become less capable of performing their function and may even die (LeDoux 2002). Now, this is a methodological question after all, since this kind of dysfunction depends on implementation properties rather than architectural properties, which would render the proposed approach inappropriate, even if it might happen that properties of particular implementations could cause abnormal behaviour (e.g. overheating of robot's parts).

From a less fundamental perspective, it could be argued that the proposed implementation is too simple, and that the effects of an emotional mechanism of the type proposed here only become relevant in the context of a reasonably complex agent architecture, e.g. one that involves a certain number of sophisticated deliberative processes competing for resources. If this turns out to be the case, then a new architecture will have to be designed.

Finally, the absence of a specific, well-defined problem to be solved might be pointed out as something that can hinder the potential of the evolutionary experiments to provide meaningful results. In other words, the virtual world may not be "realistic" enough to allow for any reasonable conclusions to be made. In this case, additional work will be necessary to create a more appropriate virtual world.

5. Conclusions

From the discussion above one can conclude that a design-based approach that incorporates results from other areas in Cognitive Science can serve as a means to investigate theories of cognitive processing as well as to provide useful architectures for intelligent systems. This is not to say that such an architecture should be adopted as a valid model of a complex cognitive system, such as an animal mind. On the contrary: we are currently at an exceedingly early stage of the architectural modelling effort to make such claims. This kind of architecture should be regarded in the same way as a scale model of an aircraft used for experiments in a wind tunnel; while it accurately models the aerodynamic surfaces of the aircraft and allows for experimenting to be done on these surfaces, it does not model any other aspects of the aircraft. And while the architecture may ultimately prove to be useful for certain kinds of applications, no exaggerated claims should be made regarding its potential.

6. References

[1] APA (1994) *DSM-IV-R: Diagnostic and Statistical Manual of Mental Disorders* (4th edition). American Psychiatric Association, 1994.

[2] Borkovec, T. (2004) Andrew Mathews: a brief history of a clinical scientist. In *Cognition, Emotion and Psychopathology: Theoretical, Empirical and Clinical Directions*. Cambridge University Press, 2004.

[3] Botelho, L. M. and Coelho, H. (2001) Machinery for Artificial Emotions. In special issue on *Grounding Emotions in Adaptive Systems* of *Cybernetics and Systems,* 2001.

[4] Damasio, A. R. (1994) *Descartes' Error: Emotion, Reason and the Human Brain*. Harper Perennial, 1994.

[5] Fox, E. (2004) Maintenance or capture of attention in anxiety-related biases? In *Cognition, Emotion and Psychopathology: Theoretical, Empirical and Clinical Directions*. Cambridge University Press, 2004.

[6] Franklin, S. (1995) *Artificial Minds*. MIT Press, 1998.

[7] Goleman, D. (1996) *Emotional Intelligence: Why it Can Matter More Than IQ*. Bantam Books, 1996.

[8] LeDoux, J. (2002) *Synaptic Self: How Our Brains Become Who We Are*. Penguin Books, 2003.

[9] LeDoux, J. (1994) Emotion, Memory and the Brain. Reprinted in the special issue on *Mysteries of the Mind* of *Scientific American*, 1997.

[10] MacLeod, C.; Campbell, L.; Rutherford, E. and Wilson, E. (2004) The causal status of anxiety-linked attentional and interpretive bias. In *Cognition, Emotion and Psychopathology: Theoretical, Empirical and Clinical Directions*. Cambridge University Press, 2004.

[11] Mathews, A. and Mackintosh, B. (1998) A Cognitive Model of Selective Processing in Anxiety. In *Cognitive Therapy and Research*, Vol. 22, No. 6, Kluwer Academic Publishing, 1998.

[12] Matthews, G.; Zeidner, M. and Roberts, R. (2003) *Emotional Intelligence: Science and Myth.* MIT Press, 2003.

[13] Michaud, F.; Robichaud, E. and Audet, J. (2001) Using Motives and Artificial Emotions for Prolonged Activity of a Group of Autonomous Robots. In *Proceedings of the AAAI Fall Symposium on Emotions*, 2001.

[14] Minsky, M. (2006) *The Emotion Machine* (draft book) available at http://www.media.mit.edu/~minsky.

[15] Minsky, M. (1987) *The Society of Mind*. MIT Press, 1987.

[16] Pinker, S. (2002) *The Blank Slate*. Viking, 2002.

[17] Pinker, S. (1997) *How the Mind Works*. W. W. Norton & Company, 1999.

[18] Richards, A. (2004) Anxiety and the resolution of ambiguity. In *Cognition, Emotion and Psychopathology: Theoretical, Empirical and Clinical Directions.* Cambridge University Press, 2004.

[19] Scheutz, M. (2004) Useful Roles of Emotion in Artificial Agents: A Case Study from Artificial Life. In *Proceedings of the AAAI 2004*, IEEE Press, 2004.

[20] Scheutz, M. and Logan, B. (2001) Affective vs. Deliberative Agent Control. In *Symposium on Emotion, Cognition and Affective Computing* at the AISB'01 Convention, 2001.

[21] Scheutz, M. and Sloman, A. (2001) Affect and Agent Control: Experiments with Simple Affective States. In *Intelligent Agent Technology: Research and Development*. Ning Zhong (ed.), World Scientific Publisher, 2001.

[22] Simon, H. A. (1967) Motivational and Emotional Controls of Cognition. Reprinted in *Models of Thought*, Yale University Press, 1979.

[23] Sloman, A. (2001) Beyond Shallow Models of Emotion. In *Cognitive Processing*, Vol. 2 No. 1, 2001.

[24] Sloman, A. (1999) How Many Separately Evolved Emotional Beasties Live Within Us? In *Emotions in Humans and Artifacts*. Trappl, R.; Petta, P. and Payr, S. (eds.), MIT Press, 2002.

[25] Sloman, A. (1998) Damasio, Descartes, Alarms and Meta-management. In *Proceedings of the International Conference on Systems, Man, and Cybernetics (SMC98)*, IEEE Press, 1998.

[26] Sloman, A.; Chrisley, R. and Scheutz, M. (2003) The Architectural Basis of Affective States and Processes. In *Who Needs Emotions? The Brain Meets the Machine*, Fellous, J.-M. & Arbib, M. (eds.), Oxford University Press, 2005.

[27] WHO (2006) *ICD-10: International Statistical Classification of Diseases and Related Health Problems* (10^{th} revision). World Health Organization, 2006.

[28] Yiend, J. (ed.) (2004) *Cognition, Emotion and Psychopathology: Theoretical, Empirical and Clinical Directions*. Cambridge University Press, 2004.