Useful Roles of Emotions in Artificial Agents: A Case Study from Artificial Life

Matthias Scheutz
Artificial Intelligence and Robotics Laboratory
Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556, USA
mscheutz@cse.nd.edu

Abstract

In this paper, we discuss the role of emotions in AI and possible ways to determine their utility for the design of artificial agents. We propose a research methodology for determining the utility of emotional control and apply it to the study of autonomous agents that compete for resources in an artificial life environment. The results show that the emotional control can improve performance in some circumstances.

Introduction

Over the last several years, emotions have received increasing attention in several AI-related fields, most prominently in human-robot/computer interaction, where emotional receptiveness (i.e., being able to perceive and interpret emotional expressions of others) and expressivity (i.e., being able to express emotions in a way that can be perceived and interpreted by others) are crucial. Within AI, believable virtual and robotic agents and human-like synthetic characters are of particular interest, with applications ranging from the entertainment industry, to training and tutoring systems (although there are also other areas of interest, e.g., affective natural language processing). The main focus in most of the employed agents is on the display of emotions (e.g., via animated facial expressions) and/or on their recognition (e.g., in speech signals).

While achieving believable emotion display and reliable emotion recognition are important goals in the context of designing virtual and robotic agents for human-computer/robot interaction, the more general question about what possible roles emotions could have in an agent architecture and in what circumstances they might be useful for the control of agents and possibly even better than other, non-emotional control mechanisms, has received very little attention.

In this paper, we outline a methodological approach, which in practice can provide at least a partial answer to the above questions. After a brief overview of past and present work on emotions in AI, we briefly describe a our approach towards studying the utility of emotions and subsequently apply it to agents in an artificial life environment. Specifically, we introduce an agent model for virtual and robotic agents that is capable of implementing emotional states, and compare the utility of its emotional control mechanism in an evolutionary survival task to other agents that do not use emotional control. The results of these simulations demonstrate the utility as well as the limits of the employed emotional control mechanisms.

Emotions–Why Bother?

Emotions have been of interest to various researchers throughout the history of AI. From very early on, architectures with emotional components have been proposed for simple and complex agents (e.g., (Toda 1962; Simon 1967; Dyer 1987), and several others; see (Pfeifer 1988) for a discussion of the early models).

Over the recent past, researchers have been particularly interested in endowing artificial agents with emotional expressivity to improve their "believability" and to make them more "life-like" (e.g., (Bates 1994; Hayes-Roth 1995; Rizzo *et al.* 1997)). Such believable virtual agents do not only find applications in the entertainment industry, but increasingly so in the realm of instruction and tutoring (e.g., (Gratch 2000; Shaw, Johnson, & Ganeshan 1999; Conati 2002)), as well as in the design of user-interfaces (e.g., (Olveres *et al.* 1998; Hudlicka & Billingsley 1999; Takeuchi, Katagiri, & Takahashi 2001)).

There is also an increasing number of examples of robotic agents that are based on emotional control (e.g., (Michaud & Audet 2001; Breazeal 2002; Arkin *et al.* 2003)), most of which are intended for human-robot interaction.

And finally, there are researchers interested in "computational models of human emotion" (e.g., (Eliott 1992; Cãnamero 1997; Marsella & Gratch 2002)), which typically are studied in simulations in artificial environments.

This list is by no means intended to give a complete overview of the present activities concerned with emotions in AI (for a more complete overview see (Picard 1997; Trappl, Petta, & Payr 2001; Pfeifer 1988)), but rather to show that an increasing number of re-

searchers are willing to investigate whether emotions *should* have a role in AI in the first place, and if so what that role might be. Before we proceed, we would like to point out that the term "emotion" has different connotations in the AI community: for some emotions are merely facial expressions, while for others emotions are intrinsically concerned with internal processes. Moreover, the same emotional terms used for human concepts (e.g., "embarrassment" or "shame") are often applied to states in artifacts without specifying how the implemented states differ from the human case. For the emotional control processes discussed in this paper we only assume a general characteristic of emotion processes (e.g., (Cãnamero 1997; Marsella & Gratch 2002)) without making any claims about the biological plausibility of the employed states or mechanisms.

Roles of Emotions in Nature

Emotional control is wide-spread in nature and seems to serve several crucial roles in animals and humans alike.

In simple organisms with limited representational capacities, for example, emotions provide a basic evaluation in terms of *hedonic values*, often causing the organism to be attracted to *what it likes* and to avoid *what it does not like*.[1] If a threat is perceivably caused by another organism, a "fear-anger" system (Berkowitz 2003) may generate "fight-or-flight" behavior (e.g., depending on an estimate of the likelihood that a fight can be won). While emotional states such as fear and anger control immediate actions (LeDoux 1996), other affective states operate on long-term behavioral dispositions (e.g., anxiety leads to increased alertness without the presence of any immediate threat).

In humans, emotions, and more generally, affect, seem to be deeply intertwined with cognition in that they can influence, bias, and direct cognitive processes and, more generally, processing strategies (e.g., negative affect, for example, can bias problem solving strategies in humans towards local, bottom-up processing, whereas positive affect can lead to global, top-down approaches (Bless, Schwarz, & Wieland 1996)). Emotions also seem to play an important role in social contexts (Cosmides & Tooby 2000), ranging from signaling emotional states (e.g., pain) through facial expressions and gestures (Ekman 1993) to perceptions of affective states that cause approval or disapproval of one's own or another agents' actions (relative to given norms), which can trigger corrective responses (e.g., guilt).

Roles of Emotions in Artificial Agents

Based on the functional roles of emotions proposed by emotion researchers for natural systems, it is worth asking whether emotions could server similar functional roles in artificial systems. Specifically, we can isolate 12 potential roles for emotions in artificial agents (beyond displaying and recognizing emotions in interactions with humans):

- *action selection* (e.g., what to do next based on the current emotional state)
- *adaptation* (e.g., short or long-term changes in behavior due to the emotional states)
- *social regulation* (e.g., communicating or exchanging information with others via emotional expressions)
- *sensory integration* (e.g., emotional filtering of data or blocking of integration)
- *alarm mechanisms* (e.g., fast reflex-like reactions in critical situations that interrupt other processes)
- *motivation* (e.g., creating motives as part of an emotional coping mechanism)
- *goal management* (e.g., creation of new goals or reprioritization of existing ones)
- *learning* (e.g., emotional evaluations as Q-values in reinforcement learning)
- *attentional focus* (e.g., selection of data to be processed based on emotional evaluation)
- *memory control* (e.g., emotional bias on memory access and retrieval as well as decay rate of memory items)
- *strategic processing* (e.g., selection of different search strategies based on overall emotional state)
- *self model* (e.g., emotions as representations of "what a situation is like for the agent")

This list is certainly not exhaustive, but provides a good starting point for systematic investigations of the utility a particular emotional control mechanisms.

A Case Study of How to Evaluate the Utility of Emotion from Artificial Life

To study the role of emotions in agent architectures and to test their utility for the control of artificial agents, we will focus on a small subset of the above list for the rest of the paper, i.e., on action selection, adaptation, and social regulation, the first three items.

We will use a general methodology for the comparison of different agents with respect to their performance in a given task and environment, which consists of four parts: (1) (emotion) concepts are analyzed and defined in terms of architectural capacities of agent architectures (Sloman 2002a), (2) agent architectures with particular (emotional) states as defined in (1) are defined for a given task together with a performance measure, (3) experiments with agents implementing these architectures (e.g., in the style of (Pollack *et al.* 1994; Hanks, Pollack, & Cohen 1993)) are carried out (either in simulations or on actual robots), and (4) the performance of the agents is measured for a predetermined set of architectural and environmental parameters. The results can then be used to relate agent performance

[1] Hedonic values seem to be at the root of various forms of reinforcement learning.

to architectural mechanisms. Moreover, by analyzing the causes for possible performance differences, it may be possible to generalize the results beyond the given task and set of environments. In the best case, general statements of the form "Mechanism X is better than mechanism Y" can be derived for whole classes of tasks and environments, where "better" is spelled out in terms of the performance ordering obtained from the experiments.

Agents and Architectures

For the experiments reported here, we used a "one-resource foraging task" for all agents in an unlimited two-dimensional environment, in which the resources appear at random within a predetermined rectangular 1800 x 1800 area at a frequency of one resource per simulation cycle (starting with 50 randomly distributed resources in the beginning). Resources contain energy (800 units), which agents need for movement and processing (agents can consume resources when they are on top of them).

All agents use a basic schema-based architecture for locomotion, which can be used for simulated and robotic agents alike (Arkin 1989). They have a perceptual system that computes directional force vectors v from their visual sensory input to resources ("resource schema") and other agents ("agent schema"). The vectors are subsequently scaled by the square of the distance to the object within their sensory range (of 300 distance units) as well as a multiplicative constant called *schema gain* depending on the type of object (g_r for resources, and g_a for agents) and summed up. The resultant vector D is sent to the motors, thus determining the direction and speed of the agent (the maximum speed is 4). For any given agent A the mapping is given by:

$$D = \sum_n g_r \cdot resource(n) + \sum_m g_a \cdot agent(m)$$

where $resource(n)$ is the vector from the position of A to the n-th resource and $agent(m)$ is the vector from the position of A to the m-th agent (not including A).

The energy expenditure for movement is the square of an agent's speed. In addition, each agent consumes one unit of energy per cycle for processing. Agents have an "energy alarm" for self-preservation that limits the energy expenditure by setting their overall speed to 1 if their energy level drops below $Energy_{crit} = 400$ (the speed will remain at 1 until the energy level is raised above the critical level again).

After $\alpha = 250$ simulation cycles agents can procreate asexually, if their energy levels are above the minimum necessary for procreation (set to 2200). The energy necessary for creating the offspring (2000) is subtracted from the parent, and an identical copy of the parent will be placed in the vicinity of the parent in the subsequent simulation cycle.

All agents compete for resources in order to survive. If two or more agents want to obtain the same resource, they have the option to "fight" for it or to "flee" from the scene. Fighting incurs a cost of 50 units per simulation cycle, whereas fleeing incurs the cost of running at a speed of 7 for about 5 to 10 cycles.

Each agent has a representation of its basic *action tendency*, which it uses to decide what to do in conflict situations. An agent's action tendency is modelled as the probability that it will fight (as opposed to flee). Agents display their action tendency and can use the action tendency displayed by their opponents to modify their own decision. We distinguish two agent kinds, *social* and *non-social* agents, depending on whether they take the opponents' action tendencies into account in a conflict. Social agents decide in the first round of a competition based on the other agents' displayed action tendency whether they will fight or flee: they will only fight if their action tendency is highest, otherwise they will flee. Asocial agents, on the other hand, simply decide their actions probabilistically based on their own action tendencies. Consequently, it is possible for two asocial agents to continue conflicts for several rounds until one finally flees (or dies), which is not possible for two social agents, as they will always determine the winner after one round.

Depending on whether an agent's action tendency can change over time or is fixed, we distinguish *adaptive* and *non-adaptive* agents. Adaptive agents change their action tendencies depending on whether they win or lose in conflicts: if they win, they lower their action tendency, thus increasing the probability that they will lose the next time; if they lose, they will raise their action tendency, thus increasing the probability that they will win the next time. This way adaptive agents implement an approximate sharing mechanism, which gives rise to *altruistic behavior* that we have shown to be beneficial elsewhere (Scheutz & Schermerhorn 2004): an agent that has not received a resource in several competitions in the past is very likely to receive one in the future, while agents that have received several resources in a row will very likely flee in the next conflict, thus giving up the resource.

There are many ways to implement such a mechanism. We use the following way to change an agent's action tendency based on its *basic action tendency* (i.e., its action tendency in the absence of any adaptation):

Definition [Adaptation Rule (AR)] Let r be the basic action tendency of agent A and let m be the current action tendency ($r = m$ if AR has never been applied). Then the $AR(m)^+$ is defined (for losses) as follows: if $m \geq r$, then $AR^+(m) = m+(1-m)/2$; if $m \leq r/2$, then $AR^+(m) = 2m$; else $AR^+(m) = r+(2m-r)(1-r)/2r$.[2] Similarly, $AR(m)^-$ is defined (for wins) as follows: if $m \geq r + (1-r)/2$, then $AR^-(m) = m - (1-m)$; if $m \leq r$, then $AR^-(m) = m/2$; else $AR^-(m) = r/2 + r(m-r)/(1-r)$.[3]

[2] This maps values in the interval $(r/2, r)$ into $(r, (1-r)/2))$.

[3] This maps $(r, (1-r)/2)$ into $(r/2, r))$.

This rule effectively keeps track of how often an individual was able to win a conflict by increasing or decreasing the action tendency relative to the basic action tendency and the current action tendency.

Finally, we introduce a third distinction between agents that have fixed schema gains, i.e., a *fixed conflict tendency* for their agent schema, and those that can adjust their agent schema gain g_a, i.e., that have a variable *variable conflict tendency*. By adjusting their conflict tendency, an agent can modify its behavioral disposition towards other agents, i.e., whether it seeks conflicts or whether it avoids them.

Emotional Agents

We can now define *emotional* agents as adaptive agents with variable conflict tendencies that adjust their agent schema gain based on their action tendency according to the following equations: $g_a = action\_tendency \cdot 100 - 50$. Hence, the emotional state of emotional agents (as defined by the current value of g_a) is directly coupled to their action tendency. If an agent's conflict tendency is positive and consequently the action tendency is above 0.5, the agent is said to be "angry", if its conflict tendency is less than zero and consequently the action tendency is below 0.5, it is said to be "fearful". Effectively, emotional agents seek conflicts with other agents when they have lost several conflicts in a row in the past (they get "angrier"), while they avoid conflicts with others if they have won conflicts (they get more "fearful"). It is this reaction to events in the environment that causes rapid, yet temporary adjustments of internal states that alter action tendencies, which warrant the attribute emotional. Specifically, the varying influence of the value of g_a on an agent's behavior can be seen as an *amplifying* or *diminishing* modification of the behavior as determined by g_f (i.e., the fixed "drive" to find and consume resources) the drives, which is typical of (some construals of) emotional states (e.g., see (Cãnamero 1997) for a similar view). The implemented states correspond to what some call "primary emotions" (e.g., (Sloman 2002b)) in that they (1) play a regulatory role, (2) are engaged automatically (by virtue of the global alarm system), and (3) alter the internal state of the agent and consequently its behavior.

It is worth mentioning that the mechanisms implemented here resemble what is presumed to be the functional organization of the emotional "fear/anger system" in many animals, given that animals are typically taken to exhibit either a "fight" or a "flee" behavior (e.g., (Berkowitz 2003)). The above model departs from biology, however, by linking the successful outcome of a conflict to an increase in fear based on the adpative sharing mechanisms–in biological systems the opposite is often true, i.e., winners will become more daring, while losers will become more cautious. However, the intent here is not to reproduce a particular biological model, but rather to investigate the utility of directly coupling the altruistic strategy implemented by adaptive agents to their conflict tendency (which, in turn, determines when and how an agent changes its emotional state). It is certainly possible to decouple the two functional components and update emotional states based on a different rule as we have done elsewhere (Scheutz 2001).

Finally, for the sake of comparison, we will also consider agents that are non-adaptive, but have a conflict tendency that is based on their fixed action tendency–they will be called *dispositional*. By default, all non-emotional, non-dispositional agents have $g_a = 0$, while $g_f = 20$ for all agents (including emotional and dispositional agents).

Experiments and Results

We examined the utility of the above strategies by comparing social and asocial emotional and dispositional agents to four types of agents with fixed conflict tendencies (i.e., non-adaptive and adaptive social and non-social agents). Specifically, we ran 16 sets of experiments, each consisting of 40 different runs for 10000 cycles each with different random initial conditions (the same 40 different initial conditions were used in all 16 sets to guarantee a fair comparison). In all runs, 25 emotional or dispositional agents of one kind and 25 non-emotional, non-dispositional agents of one kind were placed at random locations within the 1800 x 1800 resource area in the environment together with 50 randomly placed resources. The initial energy of all agents was set to 2000 and their initial action tendencies were distributed following a Gaussian distribution with spread 0.125 around 0.5. As performance measure the average numbers of survivors after 10000 cycles was used. Figures 1 and 2 show the results for emotional agents, Figures 3 and 4 for dispositional agents (the error bars indicate the 95% confidence intervals).

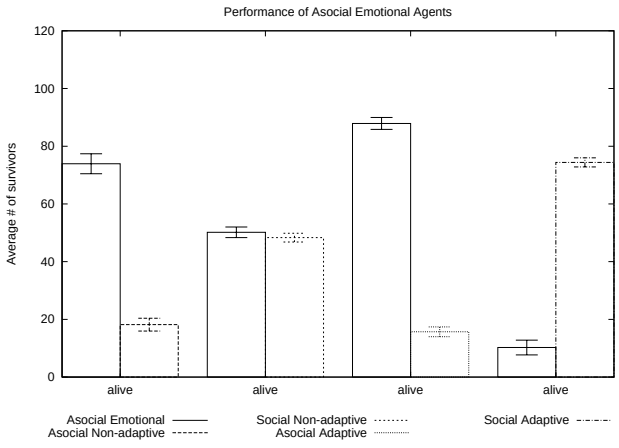

Figure 1: The average number of survivors in the experiment sets comparing asocial-emotional agents to non-emotional agents.

The results demonstrate that emotional and dispositional agents perform better than both adaptive and

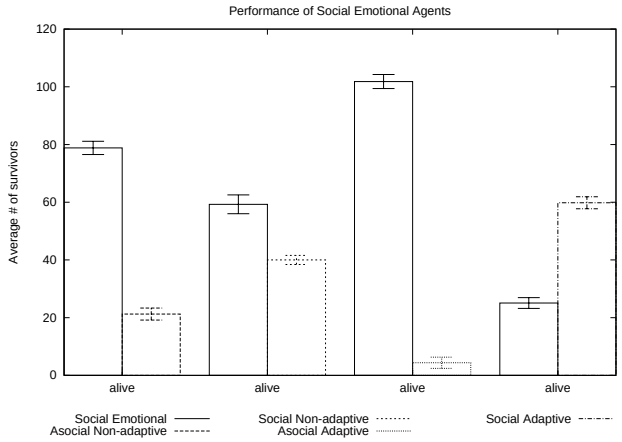

Figure 2: The average number of survivors in the experiment sets comparing social-emotional agents to non-emotional agents.

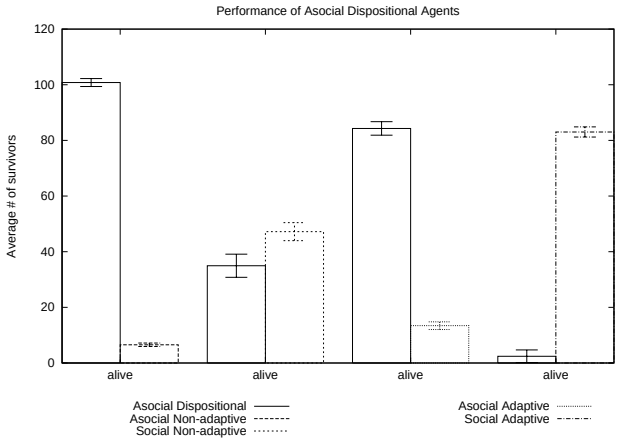

Figure 3: The average number of survivors in the experiment sets comparing asocial-dispositional agents to non-emotional agents.

non-adaptive asocial agents with $g_a = 0$. Social emotional agents show a better overall performance than non-social emotional agents, whereas social dispositional agents perform overall worse than non-social dispositional agents. Furthermore, dispositional agents perform worse than both social agent kinds with $g_a = 0$, while emotional agents only perform worse than adaptive social agents. This last result indicates that adding conflict tendencies that depend on action tendencies to social adaptive mechanisms reduces performance: loosely speaking, "fear" and "anger" do not improve performance beyond what changing one's action tendency based on the past outcomes in conflicts can achieve, but rather reduce it. It is worth noting, however, that this is only true for social agents. In the asocial case, emotional agents do fare much better, which demonstrates that changing one's conflict tendency based on one's past encounters with others is more beneficial than ignoring signals from others.

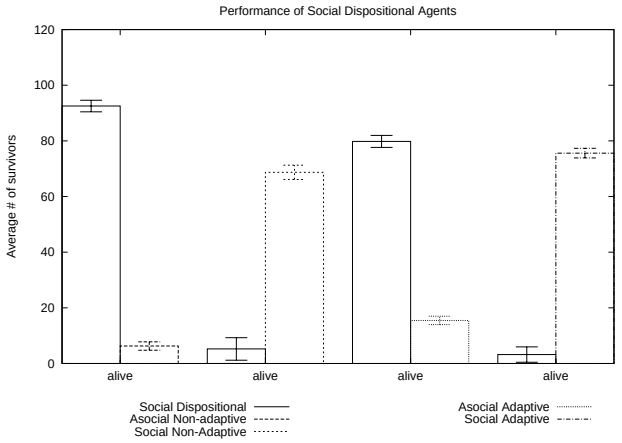

Figure 4: The average number of survivors in the experiment sets comparing social-dispositional agents to non-emotional agents.

Conclusion

In this paper, we proposed a methodology for studying possible roles of emotions in agent architectures that applies to biological organisms and artificial agents alike. We defined an emotional agent model, in which emotional processes influence the selection of actions in conflicts, the change in conflict tendencies, and consequently the distribution of resources among group members (in social emotional agents based on displaying action tendencies). All emotional agents showed high levels of performance and performed only worse than adaptive social agents. The results show that emotional control is beneficial in most circumstances in the considered environments, but that social adaptive control without influencing conflict tendencies (and thus attractive and aversive behavior)–which basically ignores other agents except in conflict–can do even better. We believe that systematic comparisons of emotional and non-emotional agents in the way suggested in this paper, will allow us, at least in part, to make objectively verifiable claims about the utility of emotional control for both biological and artificial agents in a great variety of tasks.

References

Arkin, R.; Fujita, M.; Takagi, T.; and Hasegawa, R. 2003. An ethological and emotional basis for human-robot interaction. *Robotics and Autonomous Systems* 42:3–4.

Arkin, R. C. 1989. Motor schema-based mobile robot navigation. *International Journal of Robotic Research* 8(4):92–112.

Bates, J. 1994. The role of emotion in believable agents. *Communications of the ACM* 37(7):122–125.

Berkowitz, L. 2003. Affect, aggression, and antisocial behavior. In (Davidson, Scherer, & Goldsmith 2003). 804–823.

Bless, H.; Schwarz, N.; and Wieland, R. 1996. Mood and the impact of category membership and individuating information. *European Journal of Social Psychology* 26:935–959.

Breazeal, C. L. 2002. *Designing Sociable Robots*. MIT Press.

Cãnamero, D. 1997. Modeling motivations and emotions as a basis for intelligent behavior. In Johnson, L., ed., *Proceedings of the First International Symposium on Autonomous Agents (Agents'97)*, 148–155. New York, NY: ACM.

Conati, C. 2002. Probabilistic assessment of user's emotions in educational games. *Journal of Applied Artificial Intelligence, special issue on "Merging Cognition and Affect in HCI"*.

Cosmides, L., and Tooby, J. 2000. Evolutionary psychology and the emotions. In Lewis, M., and Haviland-Jones, J. M., eds., *Handbook of Emotions*. NY: Guilford, 2nd edition. 91–115.

Davidson, R. J.; Scherer, K. R.; and Goldsmith, H. H., eds. 2003. *Handbook of Affective Sciences*. New York: Oxford University Press.

Dyer, M. G. 1987. Emotions and their computations: Three computer models. *Cognition and Emotion* 1(3):323–347.

Ekman, P. 1993. Facial expression and emotion. *American Psychologist* 48(4):384–392.

Eliott, C. 1992. *The Affective Reasoner: A process model of emotions in a multi-agent system*. Ph.D. Dissertation, Institute for the Learning Sciences,Northwestern University.

Gratch, J. 2000. Emile: Marshalling passions in training and education. In *4th International Conference on Autonomous Agents*, 325–332.

Hanks, S.; Pollack, M. E.; and Cohen, P. 1993. Benchmarks, testbeds, controlled experimentation, and the design of agent architectures. *AI Magazine* 14(4):17–42. http://www.cs.pitt.edu/ pollack/distrib/testbeds.ps.

Hayes-Roth, B. 1995. Agents on stage: Advancing the state of the art of AI. In *Proc 14th Int. Joint Conference on AI*, 967–971.

Hudlicka, E., and Billingsley, J. 1999. Affect-adaptive user interface. *Human Computer Interaction* 1:681–685.

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

Marsella, S., and Gratch, J. 2002. Modeling the influence of emotion on belief for virtual training simulations. In *Proceedings of the 11th Conference on Computer-Generated Forces and Behavior Representation*.

Michaud, F., and Audet, J. 2001. Using motives and artificial emotion for long-term activity of an autonomous robot. In *5th Autonomous Agents Conference*, 188–189. Montreal, Quebec: ACM Press.

Olveres, J.; Billingurst, M.; Savage, J.; and Holden, A. 1998. Intelligent expressive avatars. In *First Workshop on Embodied Conversational Characters*.

Pfeifer, R. 1988. Artificial intelligence models of emotion. In Hamilton, V.; Bower, G. H.; and Frijda, N. H., eds., *Cognitive Perspectives on Emotion and Motivation, volume 44 of Series D: Behavioural and Social Sciences*. Netherlands: Kluwer Academic Publishers. 287–320.

Picard, R. 1997. *Affective Computing*. Cambridge, Mass, London, England: MIT Press.

Pollack, M. E.; Joslin, D.; Nunes, A.; Ur, S.; and Ephrati, E. 1994. Experimental investigation of an agent commitment strategy. Technical Report 94-31, University of Pittsburgh. http://www.cs.pitt.edu/ pollack/distrib/tw.ps.

Rizzo, P.; Veloso, M.; Miceli, M.; and Cesta, A. 1997. Personality-driven social behaviors in believable agents. In *Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents*, 109–114.

Scheutz, M., and Schermerhorn, P. 2004. The role of signaling action tendencies in conflict resolution. *Journal of Artificial Societies and Social Simulation* 7(1).

Scheutz, M. 2001. The evolution of simple affective states in multi-agent environments. In Cañamero, D., ed., *Proceedings of AAAI Fall Symposium*, 123–128. Falmouth, MA: AAAI Press.

Shaw, E.; Johnson, W. L.; and Ganeshan, R. 1999. Pedagogical agents on the web. In *Third International Conference on Autonomous Agents*, 283–290.

Simon, H. A. 1967. Motivational and emotional controls of cognition. Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979.

Sloman, A. 2002a. Architecture-based conceptions of mind. In *Proceedings 11th International Congress of Logic, Methodology and Philosophy of Science*, 397–421. Dordrecht: Kluwer. (Synthese Library Series).

Sloman, A. 2002b. How many separately evolved emotional beasties live within us? In Trappl, R.; Petta, P.; and Payr, S., eds., *Emotions in Humans and Artifacts*. Cambridge, MA: MIT Press. 29–96.

Takeuchi, Y.; Katagiri, Y.; and Takahashi, T. 2001. Learning enhancement in web contents through interagent interaction. In Hirose, M., ed., *Eight Conference on Human-Computer Interaction (INTERACT 2001)*, 480–487.

Toda, M. 1962. The design of the fungus eater: A model of human behavior in an unsophisticated environment. *Behavioral Science* 7:164–183.

Trappl, R.; Petta, P.; and Payr, S., eds. 2001. *Emotions in Humans and Artifacts*. MIT Press.