A Framework for Evaluating Affective Control

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

Abstract

In this paper, we introduce a methodology for determining the utility of agent architectures by comparing them in an architecture framework APOC, which allows for the definition of a notion of "cost induced by an agent architecture". Using this framework, it is in particular possible to compare the performance of affective agents to that non-affective agents in a unified way taking the cost of architectural components into account.

1 Introduction

The class of affective states found in nature is comprised of many different kinds of states and processes: from mere sensations or feelings (e.g., pains), to simple homeostatic control and drives (e.g., hunger), to various kinds of motivations (e.g., to fight), to simple emotions (e.g., some forms of fear) and moods (e.g., melancholy), to all kinds of complex emotions (e.g., embarrassment), and many others.

In simple organisms with limited representational capacities, affect mainly controls behavior by providing an internal measure of "what is good and bad" for the organism [Humphrey, 1992]. The basic evaluation in terms of *hedonic values* causing the organism to be attracted to *what it likes* and to avoid *what it does not like* seems to be at the root of various forms of reinforcement learning. 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, affect seems to be deeply intertwined with cognitive processing. *Negative affect*, for example, can bias problem solving strategies in humans towards local, bottom-up processing, whereas *positive affect* seems to lead to global, top-down approaches in many cases [Schwarz,]). There is also evidence that humans often rely on *affective memory* (e.g., [Blaney, 1986]) to evaluate a situation quickly instead of performing a longer, more complex cognitive evaluation [Kahneman *et al.*, 1997] as *affective evaluations* seem to encode implicit knowledge about the likelihood of occurrence of a positive or negative future event (e.g., [Damasio, 1994]). Finally, and most importantly, affect is crucially involved in *social control* [Cosmides and 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).

However, there is no agreement among researchers in the "affective sciences" on how to define affect concepts, nor what the functional roles of affect are in agent architectures. It is, therefore, not surprising that there has been no systematic, unified approach to affect in artificial agents either [Ventura and Pinto-Ferreira, 1999; Scheutz, 2002a]), although different forms of affect have been been investigated to varying degrees ever since the beginning of AI, especially in the recent past (see [Pfeifer, 1988; Trappl *et al.*, 2001]).

While the exploration of different architectural approaches towards implementing affective control is desirable as it allows researchers to explore the space of possible designs, divergent approaches always make it difficult to compare the advantages and dis-

advanteges of the employed mechanisms, particularly, when the criteria for what it takes to implement a given property are as underspecified as in the case of affect. Worse yet, most of the proposed "affective architectures" in AI–architectures with components that can instantiate affective states–do not come with evaluation criteria that would allow for the assessment of the utility of affect (e.g., compared to other possible architectures that are appropriate for the task at hand). What is needed, however, to be able to make a general, objective statement about the utility of affect is a way to compare not only different kinds of affective architectures, but also non-affective architectures with respect to their performance (at the given task), otherwise the potential of the implemented states remains unclear.

In this paper, we introduce a methodology based on experiments with agent architectures in a unified architecture framework that allows us to evaluate the utility of affective control in an objective way (or any other property of an architecture, for that matter).

2 How can the Utility of Affect be Determined?

One difficulty with assessing the utility of affect, as already mentioned, is the intrinsic indeterminacy of affect concepts. There are several strategies of how one can overcome these problems (e.g., by restricting the concepts to be studied to clear instances of affect, e.g., fear states, where the functional role is largely well-understood). The strategy we find most useful is based on an analysis of mental concepts in terms of agent architectures, where the minimal set of requirements for a concept to be *instantiable* in an agent architecture are worked out [Sloman, 2002]–for this analysis an agent architecture framework is required [Sloman and Scheutz, 2002].

A fear state, for example, is caused by the presence of dangerous objects in the environment and changes the agent's behavioral dispositions in such as way as to make it stay away from fear elicitors. The fear state can be instantiated by a controller C, which integrates over time the frequency of occurrence of fear triggering conditions: input to C comes from an internal sensor S_f that is activated (under normal circumstances) by a fear triggering condition (e.g., the sensor outputs a unit impulse [Özbay, 2000]). C integrates these inputs over time and outputs a signal that corresponds to the intensity of "fear", hence to the degree with which the system should change its behavioral dispositions to be more alert, action-ready, etc. To be able to instantiate a fear state, the above controller C needs to be connected to the agent's effectors in a way that the positive output from C can influence and bias the agent's behavior towards avoiding or attempting to avoid dangerous objects, where the intensity with which the agent avoids or attempts to avoid these objects depends on the magnitude of the output of C (reflecting the agent's level of fear).

Once affect concepts are analyzed and defined in terms of architectural capacities of agent architectures in an architecture framework, it is possible to define architectures that implement affective states for a given task and to compare agents implemeting them to agents that implement non-affective architectures for the same task [Scheutz and Logan, 2001]. Such a comparison can be used to establish the utility of affective control relative to the non-affective architecture for the given task and investigated environmental conditions. In particular, the absolute utility of affective control for the given task can be obtained if the non-affective architecture is "minimal", where what "minimal" means is fleshed out in terms of a notion of "cost induced by an architecture", which we will define below (alternatively, it could be measured in terms of the smallest number of states of Turing machine that implements it along the lines of algorithmic information theory, e.g., [Chaitin, 1992]).

For the comparison of different architectures, a *performance measure* is required, which could be task-dependent. For example, in evolutionary studies in an artificial life environment the performance measure might be the average number of surviving agents after a fixed number of simulation steps. For a robot that needs to detect affect expression in human faces, on the other hand, it might be a combination of the number of faces recognized and the number of affective features detected properly (e.g., an angry face).

It is important to note that the utility of an architecture is essentially based on the employed performance measure. Hence it is possible that the same two architectures have different utilities for different performance measures. Especially, in the case of affect, it is, therefore, crucial to settle on a notion of utility that reflects our intuitions about why affect should or could be advantegeous (e.g., when resources are scarce, fast decisions are required, information is incomplete or unreliable, etc. [Scheutz, 2001b]). For example, comparing a "rational agent" playing chess based on minimax search in terms of number of games won to an affective agent that selects

partly suboptimal moves based on emotional states (such as "frustration", "anger", "disappointment", "pleasure", etc.) is not a performance measures that will be particularly useful in determining the utility of affect. For one, because emotional states might be beneficial not because they give rise to better absolute performance, but rather yield better *relative performance*. I.e., if it is true that (some of) the roles of affect are to do with providing quick and efficient means to reach decisions of importance to an organism that are by and large *good decisions* (as many psychologists claim), then the performance of an affective agent needs to be evaluated relative to *the cost* involved in reaching the decision. Hence, it seems that for affect the *relative performance-cost tradeoff* is the critical measure to evaluate its utility. In the case of the chess playing agents, this means that the number of games won by the rational agent has to be related to the computational cost of carrying out minimax search to be able to compare it to the number of games won by the affective agent (also taken relative to the computational cost of the affective decision processes).

Once the performance measure is defined, experiments with two kinds of agents, one implementing affective, the other implementing non-affective architectures, can be carried out to determine their actual performance. The number of experiments and the variation over initial conditions to ensure that the results are not dependent on particular favorable conditions will vary dependent on the given task and kinds of agents (i.e., virtual or robotic). In simulation experiments, it is usually possible to average over a large number of initial conditions, whereas in robotic experiments the number of variations will be confined to what can be achieved in a reasonable amount of time for practical reasons. In addition to initial conditions, a set of architectural parameters will typically be specified which are also systematically varied. For example, in the case of the "fear controller" the different control parameters of the control circuit are subject to variation in order to determine which parameter settings maximize the performance of the "fearful agents".

The result of the experiments constitutes a *performance space*, based on the set of parameters that were open to variations (i.e., the *architecture space*). By comparing the performance spaces of agents with different agent architectures it is then possible to determine the absolute or relative utility of affective control for the whole range of parameters. In the case, where all parameters relevant to the task have been varied in their whole ranges, the outcome will be about the utility of affective control for the *task per se* (without any restricting conditions).

All resulting performance spaces are then compared, in particular, with respect to the agents' *(relative) performance-cost tradeoffs*, i.e., their performance taken relative to the (computational) cost necessary to maintain and run the instantiated architecture. As already mentioned, the relative comparison is critical for affective agents as they might not do better than non-affective agents on the given task in absolute terms.

The following section provides more details on the concepts involved in this methodology.

3 The Notion of "Cost Induced by an Agent Architecture"

Over the last years, we have developed the *agent architecture framework* APOC ("Activating-Processing-Observing-Components") [Andronache and Scheutz, 2002; 2003a; 2003b], which provides a unified framework into which other architectures can be translated. APOC consists of heterogeneous computational units (based on [Scheutz, 2001a]) called *components* that can be connected via four link types to form an agent architecture. The four link types cover important basic interaction types among components in agent architectures: the *activation link* (A-link) allows components to communicate with other components; the *observation link* (O-link) allows components to observe the state of other components; the *process control link* (P-link) enables components to influence the computation taking place in other components, and finally the *component link* (C-link) allows a component to instantiate other components and connect to them via the other three links.

Components can vary with respect to their complexity and the level of abstraction at which they are defined. They could be as simple as a connectionist unit or as complex as a full-fledged condition-action rule interpreter. APOC can be used as an analysis tool for the evaluation of architectures, since it can express any agent architectures in a unified way (e.g., cognitive architectures such as SOAR, ACT-R, and others, as well as behavior-based architectures such as subsumption, motor schemas, situated automata, etc.).

Most importantly, it introduces a novel idea that is essential for the study of architectural trade-offs: a notion of *cost induced by*

an architecture, which is defined in terms of the cost associated with *structures*, *processes*, and *actions on the architecture*. This notion is different from other notions of cost that have been defined for processes in terms of process algebras or π-calculus [Milner, 1993; Eberbach, 2001].[1]

Structural costs are those that are incurred as a result of merely having a certain component or link instantiated. They can be thought of as maintenance costs that are associated with any work that needs to be done to keep the object up to date. *Process costs* are those associated with running processes; they include computational costs, and possibly the costs of I/O and other such operations. Typically process costs will be proportional to the complexity of the computation performed by the process. Finally, *action costs* are those associated with primitive operations on the architecture (such as instantiating a new component or link, or interrupting a process). Each action has a fixed cost, making the computation of action costs a simple matter of assessing the associated cost whenever the action is executed. The notion of cost induced by an architecture is then inductively defined in terms of these three basic cost types.

Using the notion of cost induced by an architecture, the notion of *performance-cost-tradeoff* $PCT(P,A,T,E)$ for an agent architecture A and a task T in an environment E can be defined as P/C, where P is the given performance measure for T and C is the cost of A for T in E.[2] Mathematically, performance-cost trade-offs are orders, and can thus form the basis of the comparison of agent architectures: given an order $>_P$ defined on P, an architecture A is said to be *better* than an architecture B with respect to T, E, and P, if $PCT(P,A,T,E) >_P PCT(P,B,T,E)$.

Furthermore, given an architectural parameter λ of an architecture A that can be varied and its set of possible values Λ, PCT can be used to define an order on the *space of architectures* $A_{\lambda\in\Lambda}$. An architecture space $A_{\lambda\in\Lambda}$ is said to be *relatively better* than an architecture space $B_{\lambda\in\Lambda}$ (with respect to T, E, and P), if there exists an architecture $A \in A_{\lambda\in\Lambda}$ which is better than every architecture $B \in B_{\lambda\in\Lambda}$. $A_{\lambda\in\Lambda}$ is said to be *absolutely better* than an architecture space $B_{\lambda\in\Lambda}$ (with respect to T, E, and P), if $A_{\lambda=c} \in A_{\lambda\in\Lambda}$ is better than $B_{\lambda=c} \in B_{\lambda\in\Lambda}$ for every $c \in \Lambda$. The former measure is particularly important for evolutionary settings as a relatively better architecture space will probably be favored by evolutionary methods (i.e., evolutionary search is likely to find the best architectures in the relatively better space). The latter measure is particularly important for architecture design, since architectures from absolutely better architecture spaces are always to be preferred (for the given task, environment and performance measure). It should be noted that all of the above order notions can be directly extended to sets of tasks, environments, and performances measures.

4 Discussion

We have applied the above methodology of studying affect in artificial agents in several preliminary investigations with mostly simple agents. Our results show, for example, that *affective action selection* can be very effective in the competition for resources in hostile multiagent environments [Scheutz, 2000; Scheutz *et al.*, 2000; Scheutz, under review]. Affective control mechanisms performed much better in a variety of foraging, survival, and object collection tasks in environments with little to no structure than agents with much more sophisticated deliberative control systems (including A^*_ε planning [Pearl, 1982], plan executing methods with error feedback, and goal management mechanisms) if the "cost of deliberation" is taken into account [Scheutz and Logan, 2001; Scheutz and Schermerhorn, 2002; 2003]. Furthermore, we found that simple affective states (such as motivational "hunger" and "thirst" states [Scheutz and Sloman, 2001], or emotional states like "fear" and "aggression" [Scheutz, 2001b]) are likely to evolve in a variety of competitive multiagent environments. Finally, in studies of the potential of *affect expression and recognition for social control* we found that affect can have a beneficial regulatory effect in social groups [Scheutz, 2002b] and lead to superior conflict resolution strategies [Scheutz and Schermerhorn, forthcoming].

Most of these previous results, however, did not attempt a detailed break-down of the costs induced by the architectures, but rather assumed an overall global cost. We are planning on repeating several of these experiments with a much more detailed analysis in terms

[1] It is not trivial to define a notion of cost for agent architectures because cost is typically (i.e., in complexity theory) not assessed with respect to ongoing processes, where inputs are not known *apriori*, but are changing based on the interaction of the agent with its environment, which are impossible to predict [Wegner, 1997].

[2] Note that performance measures can be numeric, but may also consist of non-numeric entities so long as an order $>_P$ and a quotient P/C (for the involved notion of cost) can be defined.

of structural, process, and action costs. This should give us a better picture of what the overheads of affective processing are as well as what kinds of mechanisms minimize the processing cost while still being able to implement affective control processes. Moreover, it should be possible to determine a "cost tradeoff" between structural and processing costs (e.g., the difference between an emotional control mechanism that is based on a perceptual component that suppresses active behaviors and carries out emergency responses when emotion elicitors are perceived, compared to a modulating component that is active all the time and simply rearranges overall behavioral dispositions). Eventually, we would like to map out the whole space of the proposed schema-based architecture for a large number of different environments for the given foraging task to be able to generalize our previous results to cover all environmental situations, in which agents can successfully survive in the long run. Such a generalized result would provide a yard-stick against which other (especially non-affective) approaches to solving the foraging/survial task could be compared and would eventually allow us to justify claims such as "affect is beneficial for tasks of kind X" that make general statements about the utility of affective control.

References

[Andronache and Scheutz, 2002] Virgil Andronache and Matthias Scheutz. Contention scheduling: A viable action-selection mechanism for robotics? In Sumali Conlon, editor, *Proceedings of the Thirteenth Midwest Artificial Intelligence and Cognitive Science Conference, MAICS 2002*, pages 122–129, Chicago, Illinois, April 2002. AAAI Press.

[Andronache and Scheutz, 2003a] V. Andronache and M. Scheutz. APOC - a framework for complex agents. In *Proceedings of AAAI Spring Symposium 2003*. AAAI Press, 2003.

[Andronache and Scheutz, 2003b] V. Andronache and M. Scheutz. Growing agents - an investigation of architectural mechanisms for the specification of 'developing' agent architectures. In *Proceedings of FLAIRS 2003*, pages 2–6. AAAI Press, 2003.

[Berkowitz, 2003] Leonard Berkowitz. Affect, aggression, and antisocial behavior. In [Davidson *et al.*, 2003], pages 804–823. 2003.

[Blaney, 1986] P. H. Blaney. Affect and memory: A review. *Psychological Bulletin*, 99(2):229–246, 1986.

[Chaitin, 1992] G. J. Chaitin. *Algorithmic Information Theory*. Cambridge University Press, fourth printing edition, 1992.

[Cosmides and Tooby, 2000] Leda Cosmides and John Tooby. Evolutionary psychology and the emotions. In M. Lewis and J. M. Haviland-Jones, editors, *Handbook of Emotions*, pages 91–115. Guilford, NY, 2nd edition, 2000.

[Damasio, 1994] A. R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Gosset/Putnam Press, New York, NY, 1994.

[Davidson *et al.*, 2003] Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith, editors. *Handbook of Affective Sciences*. Oxford University Press, New York, 2003.

[Eberbach, 2001] E. Eberbach. Evolutionary computation as a multi-agent search: A $-calculus perspective for its completeness and optimality. In *Proceedings of Congress on Evolutionary Computation CEC'2001*, pages 823–830, Seoul, Korea, 2001.

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

[Humphrey, 1992] Nicholas Humphrey. *A History Of The Mind*. Chatto and Windus, London, 1992.

[Kahneman *et al.*, 1997] D. Kahneman, P.P. Wakker, and R. Sarin. Back to bentham? explorations of experienced utility. *Quarterly Journal of Economics*, 112:375–405, 1997.

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

[Milner, 1993] Robin Milner. Elements of interaction: Turing award lecture. *Communications of the ACM*, 36(1):78–89, 1993.

[Özbay, 2000] Hitay Özbay. *Introduction to feedback control theory*. CRC Press, London, 2000.

[Pearl, 1982] J. Pearl. A^*_ε—an algorithm using search effort estimates. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 4, pages 392–399, 1982.

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

[Scheutz and Logan, 2001] Matthias Scheutz and Brian Logan. Affective versus deliberative agent control. In Simmon Colton, editor, *Proceedings of the AISB'01 Symposium on Emotion, Cognition and Affective Computing*, pages 1–10, York, 2001. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

[Scheutz and Schermerhorn, 2002] Matthias Scheutz and Paul Schermerhorn. Steps towards a theory of possible trajectories from reactive to deliberative control systems. In Russell Standish, editor, *Proceedings of the 8th Conference of Artificial Life*. MIT Press, 2002.

[Scheutz and Schermerhorn, 2003] Matthias Scheutz and Paul Schermerhorn. Many is more but not too many: Dimensions of cooperation of agents with and without predictive capabilities. In *Proceedings of IEEE/WIC IAT-2003*. IEEE Computer Society Press, 2003.

[Scheutz and Schermerhorn, forthcoming] Matthias Scheutz and Paul Schermerhorn. The role of signaling action tendencies in conflict resolution. *Journal of Artificial Societies and Social Simulation*, forthcoming.

[Scheutz and Sloman, 2001] Matthias Scheutz and Aaron Sloman. Affect and agent control: Experiments with simple affective states. In Ning Zhong, Jiming Liu, Setsuo Ohsuga, and Jeffrey Bradshaw, editors, *Intelligent Agent Technology: Research and Development*, pages 200–209. World Scientific Publisher, New Jersey, 2001.

[Scheutz *et al.*, 2000] Matthias Scheutz, Aaron Sloman, and Brian Logan. Emotional states and realistic agent behaviour. In Philippe Geril, editor, *Proceedings of GameOn 2000, Imperial College London*, pages 81–88, Delft, 2000. Society for Computer Simulation.

[Scheutz, 2000] Matthias Scheutz. Surviving in a hostile multiagent environment: How simple affective states can aid in the competition for resources. In Howard J. Hamilton, editor, *Advances in Artificial Intelligence, 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000, Montréal, Quebec, Canada, May 14-17, 2000, Proceedings*, volume 1822 of *Lecture Notes in Computer Science*, pages 389–399. Springer, 2000.

[Scheutz, 2001a] Matthias Scheutz. Ethology and functionalism: Behavioral descriptions as the link between physical and functional descriptions. *Evolution and Cognition*, 7(2):164–171, 2001.

[Scheutz, 2001b] Matthias Scheutz. The evolution of simple affective states in multiagent environments. In Dolores Cañamero, editor, *Proceedings of AAAI Fall Symposium*, pages 123–128, Falmouth, MA, 2001. AAAI Press.

[Scheutz, 2002a] Matthias Scheutz. Agents with or without emotions? In Rosina Weber, editor, *Proceedings of the 15th International FLAIRS Conference*, pages 89–94. AAAI Press, 2002.

[Scheutz, 2002b] Matthias Scheutz. The evolution of affective states and social control. In Charlotte K. Hemelrijk, editor, *Proceedings of International Workshop on Self-Organisation and Evolution of Social Behaviour*, Monte Verità, Switzerland, 2002.

[Scheutz, under review] Matthias Scheutz. Schema-based architectural approach towards implementing affective states in autonomous agents. under review.

[Schwarz,] N. Schwarz. Feelings as information: Informational and motivational functions of affective states.

[Sloman and Scheutz, 2002] Aaron Sloman and Matthias Scheutz. A framework for comparing agent architectures. In *UK Workshop on Computational Intelligence*, 2002.

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

[Trappl *et al.*, 2001] R. Trappl, P. Petta, and S. Payr, editors. MIT Press, 2001.

[Ventura and Pinto-Ferreira, 1999] R. Ventura and C. Pinto-Ferreira. Emotion-based agents: Three approaches to implementation. In J. Velásquez, editor, *Third International Conference on Autonomous Agents, Workshop on Emotion-Based Agent Architectures*, pages 121–129, Seattle, USA, 1999.

[Wegner, 1997] P. Wegner. Why interaction is more powerful than algorithms. *Communications of the ACM*, 40(5):80–91, 1997.