Architectural Roles of Affect and How to Evaluate Them in Artificial Agents

Matthias Scheutz
Human-Robot Interaction Laboratory
Department of Computer Science
Tufts University, Medford, MA 02155, USA
mscheutz@cs.tufts.edu

Abstract

Is it possible to design affective artificial agents and if so, why should we care? This paper addresses both questions by laying out a program for systematically defining and evaluating possible functional roles of affective states in architectures for virtual and robotic artificial agents. It provides functional and architectural characterizations for simple and complex affective states, discusses possible interactions between affective and non-affective processes, and proposes an experimental evaluation framework that allows for the rigorous quantification of the utility of architectural components (for affective and non-affective agents alike). In doing so, it also provides a brief overview of past findings about the utility of affect mechanisms for artificial agents that were obtained following the proposed methodology.

Keywords: utility of affect, affective agent architectures, evaluation of affective agents, virtual and robotic agents, functional roles of affect

Introduction

Over the last fifteen years, researchers have started to investigate different forms and roles of affect in virtual agents and robots. Their efforts are, at least in part, based on the recognition that affective states like pleasure, happiness, elation, admiration, anxiety, remorse, disgust, anger, and many others are involved in many cognitive processes in humans and animals, and that, as a result, affective processes might have important functions in cognitive architectures which might benefit artificial agents. There is ample evidence from psychology (Frijda, 1986; Izard, 1993; Scherer, Schorr, & Johnstone, 2001), neuroscience (Damasio, 1994; LeDoux & Fellous, 1995; Panksepp, 2000; Hamm, Schupp, & Weike, 2003), and ethology (Lorenz & Leyhausen, 1973; Lorenz, 1981; McFarland, 1981) that affective processes are, among other functions, involved in (1) the initiation, selection, regulation and coordination of behavior, (2) the management of motivation and goals, (3) the formation of memories and memory recall, (4) attentional control, (5) different forms of associative and reinforcement learning, (6) social signaling and reacting to signals of other animals.

In simple organisms with limited cognitive and representational capacities, affect seems to control behaviors mainly by providing an internal measure of “what is good and bad” for the organism (e.g., 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* (e.g., Gray, 1990) forms the basis of the organism’s behaviors. If another organism poses a perceivable threat, a *fear-anger system* (Berkowitz, 2003) may generate *fight-or-flight* behavior. And while emotional states such as fear and anger control immediate actions (LeDoux, 1996), other affective states may operate on longer term behavioral dispositions (e.g., anxiety caused by repeated triggering of fear leads to increased alertness without the presence of any immediate threat).

In humans, affect is deeply ingrained in the cognitive architecture, biasing and influencing attentional mechanisms (such as interrupting and distracting the current processing; see Derryberry & Tucker, 1994, or broadening the attentional focus, see Fredrickson, 1998). Positive and negative affect often influence problem solving strategies, with negative affect causing local, bottom-up processing, while positive affect tends to cause global, top-down approaches in many cases (Bless, Schwarz, & Wieland, 1996; Schwarz, 1990). Humans also seem to often rely on affective memory (e.g., Blaney, 1986; Bower & Cohen, 1982) to evaluate a situation quickly instead of performing a longer, more complex cognitive evaluation (Kahneman, Wakker, & Sarin, 1997), which suggests that affective evaluations might encode implicit knowledge about the likelihood of occurrence of positive or negative future events (e.g., Damasio, 1994; Clore, Gasper, & Conway, 2001). Finally, affect is crucially involved in social coordination (Frijda, 2000; 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., feeling guilty).

All this evidence for the crucial role of affect in human and animal cognition seems to point to the potential affective control might have for artificial agents. However, outside of the human-computer interface community which, has embraced affective computing for some time (Picard, 1997), architectural affect mechanisms have not yet entered main-stream AI and robotics research (even though there have been several notable attempts over the years, e.g., Murphy, Lisetti, Tardif, Irish, & Gage, 2002; Breazeal, 2002; Arkin, Fujita, Takagi, & Hasegawa, 2003). Part of the problem might be that it is challenging to clearly define affect concepts and thus propose specific mechanisms that implement them. Moreover, there is often little agreement among researchers working on affective agents what the right algorithms and state update rules are for different types of affective states. Finally, there is a vexing general issue that has not been addressed sufficiently and calls into question the utility of pursuing affective agents: why should we consider potentially ill-defined affect mechanisms with unclear performance if we already know operational algorithms, possibly even the provably best ones, for a given problem? Clearly, one cannot beat the optimal solution.

While there is no simple answer to this challenge, it is possible to address it using a research strategy that grounds affect concepts in architectural terms and proceeds with a systematic investigation of affective agents in a clearly specified task under varying environmental conditions: “The proposed research strategy then is to start with a notion of affective state, which is applicable to natural systems, determine/define its function in a particular agent architecture and subsequently try to explore the properties of this state for concrete agents in different environments with the goal of extending the notion to more complex cases. This includes investigating ways in which slight changes in environments can change the trade-offs between design options for the architecture and hence for the functional role of the affective state” (Scheutz, 2001).

In this paper, we will lay out in more detail this strategy of defining architectural mechanisms for and evaluating the utility of affective states in artificial agents. We start with a discussion of possible architectural roles of affect in artificial agents, ranging from simple motivations and emotions as they are found in many animals, to more complex human-like motivations and emotions. We also briefly address the possible interplay between affective and rational processes. Then, we propose a framework that allows for the rigorous evaluation of affective states across architectures, tasks, and environments and the comparison of affective vs. non-affective control mechanisms. The subsequent discussion section then reports on some of our own findings over the last decade which we obtained using the proposed framework and sketches a useful direction for refining the experiments to obtain more general results. Finally, the conclusion section provides a brief summary of the proposed research program.

Functional and Architectural Aspects of Affect

A serious challenge in exploring architectural mechanisms of affect is the sheer complexity of the natural affect domain, ranging from very simple control states with little to no representational and processing requirements, to complex deliberative states with extensive representational and processing demands. In fact, "affect", like many folk notions, is a cluster concept that does not allow for the specification of necessary and sufficient conditions for something to fall under it. Yet, "there are subclasses of affective states that do share common properties or relationships, such as their functional role in a particular part of an agent architecture or the architectural requirements for those states." (Scheutz & Sloman, 2001). In the following, we will illustrate the differences in functional roles and architectural requirements for simple as well as complex motivations and emotions.

Simple Motivations and Emotions

Functionally, simple motivations can be caused by the disparity between an agent's *desire state* and the perceived state of the internal ("body") or external ("world") environment. As such, simple motivations are targeted at reducing the disparity and can thus themselves be causes for actions that attempt to change the internal or external environment in such a way as to make it agree with the agents' desired states (Sloman, Chrisley, & Scheutz, 2005). Architecturally, different types of proportional controllers can be used to implement disparity reducing mechanisms. A "hunger state", for example, can be implemented using a proportional controller connected to an internal energy sensor that measures the current energy level. The difference between the actual and the desired energy level (given by the set point of the controller) then generates a control signal that can be used for action selection. The magnitude of the difference can be used to model the intensity of the motivation and is, itself, a measure of the urgency with which the system requires energy.

Simple emotions are functionally similar to simple motivations except that they themselves typically are the states that the agent does or does not desire. 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. Architecturally, the fear state can be instantiated by a controller, which integrates over time the frequency of occurrence of fear triggering conditions. Input to the controller comes from an internal sensor that is activated (under normal circumstances) by a fear triggering condition. The controller then 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 needs to be connected to the agent's effectors in a way that the output from the controller 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 controller's output, thus reflecting the agent's level of fear.

Note that both simple motivations and emotions, as described above, can be implemented via simple feedback control circuits without any representational requirements.[1] There are many examples of virtual and robotic agents that use these types of motivational and emotional control

[1] It is important to add that there are variants of both types of states with different functional properties, e.g., other forms of hunger such as "gusto" or of anger such as "anger at a thought", both of which might require representational mechanisms.

mechanisms, although they often differ with respect to the particular functional details of the control loop (e.g., Cãnamero, 1997; Gadanho, 2003 for simulated and Nourbakhsh et al., 1999; Michaud & Audet, 2001; Breazeal, 2002; Murphy et al., 2002; and Arkin et al., 2003 for robotic agents). Additional differences among the various proposals can be found in the way these controllers are implemented and whether (and how) the authors justify their being "affective" or "emotional".

Complex Motivations and Emotions

Contrast the above simple hunger and fear states with complex motivational and emotional states such as "desiring to win a grant" or "worrying about whether the grant proposal can be completed in time". On the face of it, the former is similar in to the simple motivational hunger state in that it is goal directed and intended to reduce a disparity between the current state ("no grant") and a future state ("having won a grant"). Analogously, the latter is similar to the simple emotional fear state in that it is the worrying state that the agent does not desire and that it is the worry that can drive motivations to take a course of action. At the same time, these two states are very rich in internal representational and cognitive structure, different from the simple states, and can be caused by a great variety of factors both external and internal to the agent (e.g., becoming motivated to write a grant to support one's research group). Moreover, complex affective states have many different kinds of "intensity aspects" (such as "urgency", "importance", "significance", etc.) and different kinds valenced states associated and co-occurring with them (e.g., feelings of pleasure from the anticipated grant award and displeasure from having to go through the whole grant writing phase). Consequently, complex motivations and emotions will require complex representations, which can be mutually dependent on each other (thus forming recursive and circular data-structures). Processing then involves the update of all the components of such states, where the update of each component can range from simple operations (such as "increase the urgency linearly over time") to complex deliberative processes (such as "determine the likely outcome of a coping strategy that will change the hedonic value of the emotion based on current evidence and adjust the urgency of the motivation accompanying the emotion if the utility turns out to be lower than previously expected, possibly dropping the motivation", which could lead to sequences of adjustments and updates of other emotional and motivational states).

Specifically, a complex motivation such as "desiring to win a grant" may include any of the following components and possibly others (based on the analysis in Beaudoin & Sloman, 1993):

1. a proposition denoting a possible state of affairs, which can be true or false (e.g., *the grant has been awarded*)
2. a motivational attitude towards (1) (e.g., *make true*)
3. a value representing the intensity with which (1) is desired (e.g., *very high*)
4. a belief about (1) which together with (2) and (3) disposes the agent to act on it (e.g., *(1) is false*)
5. a value representing the importance of (2) with respect to various factors such as beliefs, norms, standards, and other goals (e.g., *(1) is crucial to research career*)
6. a measure of the urgency to act on (1) given the current situation (e.g., *do not wait*)
7. a value representing the strength with which the agent is disposed to act on (1) (based on (2) through (5)) (e.g., *very high*)
8. a plan or set of plans for achieving (1) (e.g., *pick an open problem*, *find a solution*, *formulate*

a research approach, review the literature, ...)

9. a commitment status (e.g., *adopted*)
10. management information to determine when action should begin or be resumed (e.g., *nearly completed*)
11. status information (e.g., *current*)

Similarly, complex emotions such "worrying about whether the grant proposal can be completed in time" may include any of the following (based on (Ortony, Clore, & Collins, 1988, Wehrle & Scherer, 2001) and others) and possibly more:

1. an elicitor (e.g., *the grant proposal*)
2. an eliciting condition (e.g., *the possibility of (1) not being completed by the deadline*)
3. criteria for the evaluation of (2) based on various factors such as beliefs, goals, norms, standards, tastes, attitudes, etc. (e.g., *completing (1) by the deadline is crucial to research career*)
4. an evaluation of (2) in terms of (3) (e.g., *(2) is undesirable*)
5. possible causes for (2) (e.g., *deadline approaching rapidly, work progressing too slowly, etc.*)
6. a hedonic attitude towards (2) (e.g., *displeasure*)
7. a measure of the urgency to act on (1) given (2) (e.g., *urgent*)
8. a set of strategy to cope with (2) (e.g., *cancel meetings, focus attention on (1), etc.*)
9. a set of motivations to be instantiated based on (8) (e.g., *being able to continue one's research, being able to fund students, etc.*)
10. a set of emotions to be instantiated based on (4) through (8) (e.g., *distress*)
11. the selected motivation (if any) based on (4) through (8) (e.g., *being able to continue research*)

There are very few examples of implemented systems that involve complex motivations and emotions (e.g., Wright, Sloman, & Beaudoin, 1996; Dyer, 1987, Mueller, 1998; and Gratch & Marsella, 2004 for simulated agents and Scheutz, Schermerhorn, Kramer, & Middendorff, 2006 for robots). To some extent this may be due to the architectural requirements (in terms of representational features and processing mechanisms) of these states, but more importantly we believe it reflects the fact that it is still unclear what roles these kinds of states could play in an agent architecture.

Functional Roles of Affect

A first step to address this problem would be to start with the functional roles of affective states in natural systems and ask whether they could serve similar functional roles in artificial systems. Note that there are two important assumptions underlying this question: (1) that affect *can have functional roles* in agent architectures, and (2) that these functional roles are *independent of the particular physical makeup* of the agent.

Most researchers in the affective sciences will agree on (1) (even though there are many examples of the effects of *dysfunctional affect* as well). Views diverge on the role of the visceral processes involved in and accompanying many affective states. If the exact physical nature of visceral effects does not play a causal role in the functioning of affect processes (e.g., if simulated hormonal systems could be used to achieve the same effects, e.g., Cãnamero, 1997 and Allen, 2001), then artificial agents will be able to instantiate affect processes if they have the right architectural prerequisites (and the right kinds of environmental circumstances for instantiation of affective states obtain for those affective states that intrinsically depend on external factors).

On the other hand, if the *particular visceral processes* (such as the secretion of particular hormones, changes of particular neurotransmitters, etc.) are taken to be *essential* to or *constitutive* of affect, artificial agents will, *by definition*, be incapable of instantiating affective states so construed (e.g., see Hayes-Roth, Ball, Picard, Lisetti, & Stern, 1998 and cp. to views some philosophers have voiced about consciousness or qualitative states, e.g., Searle, 1992). However, even assuming this position, it is still possible to replicate functional aspects of affect in artificial agents, i.e., the same kinds of *control processes* (as implemented in neural activity in animals) which are, also by definition, independent of the physical make-up of an agent and may be sufficient for AI purposes (e.g., for an artificial agent to be able to perform a particular task).

Hence, one's views on the relationship between visceral processes and affective states do not preclude the specification of their functional roles in an agent architecture. Similarly, regardless of what stance one wants to take on the qualitative nature of affect (i.e., one's answer to the question "what it is like to experience state X?"), the functional aspects of affect in the context of an agent's control system can be independently considered. We have compiled a non-exhaustive list of twelve potential roles of affect in architectures for artificial agents (Scheutz, 2004c):

1. *alarm mechanisms* (e.g., fast reflex-like reactions in critical situations that interrupt other processes)
2. *action selection* (e.g., what to do next based on the current affective state)
3. *adaptation* (e.g., short or long-term changes in behavior due to the affective states)
4. *social regulation* (e.g., using affective signals to achieve social effects)
5. *learning* (e.g., affective evaluations as Q-values in reinforcement learning)
6. *motivation* (e.g., creating motives as part of an emotional coping mechanism)
7. *goal management* (e.g., creation of new goals or reprioritization of existing ones)
8. *information integration* (e.g., affective filtering of data from various information channels or blocking of such integration)
9. *attentional focus* (e.g., selection of data to be processed based on affective evaluation)
10. *memory control* (e.g., affective bias on memory access and retrieval as well as decay rate of memory items)
11. *strategic processing* (e.g., selection of different search strategies based on overall affective state)
12. *self model* (e.g., affect as representations of "what a situation is like for the agent")

While this list is not intended to be exhaustive, it does provide a context for locating past work and future research on architectural aspects of affect. Most work to date, for example, has focused on the first few roles, in particular, attention has been given to affective or emotional action selection, both in simulated agents (e.g., Gadanho, 2003) and robotic agents (e.g., Arkin et al., 2003), for obvious reasons. Similarly, quite a bit of work has investigated the utility of (e)valuations that are internally generated and reflect some aspect of the internal environment (rather than the external environment) for reinforcement learning, even though most of these investigations do not call these (e)valuations “affective”. Yet, surprisingly little work has focused on investigating roles 6 through 11, which focus on the interplay between affect and deliberative processes (e.g., reasoning, planning, etc.), although there are some notable exceptions (e.g., El-Nasr, Yen, & Ioerger, 2000; Eliott, 1992; Gratch & Marsella, 2004). Finally, role 12 might turn out to be of importance for reflective systems (e.g., as described in Sloman & Chrisley, 2003, etc.; also cp. *“Conscious” Mattie*, see Franklin, Kelemen, & McCauley, 1998) that have various kinds of representations and models of themselves, which they can use for processing.

Moods and Rationality

One important question for affective agents is whether and how affective and non-affective—in particular, rational—processes can co-exist and interact in the same agent architecture; for according to a frequently held, but in our view false belief, affective and rational decision mechanisms are *mutually exclusive*. To see that this is not so, observe that rational and affective processes can work in parallel and mutually support each other. For example, an agent may have the option to decide which mechanism to prefer on a case-by-case basis based on the time it has to make a decision: if the agent has sufficient time and computational resources, it might determine its actions by a purely decision-theoretic process involving utility measures and likely action outcomes to determine the best action, i.e., the one maximizing the expected utility; alternatively, when time and processing resources are scarce, the agent might rely on affective decision processes which may not select the best action, but an action “good enough” for the agent’s current purposes. More specifically, a perfectly rational agent with perfect information can make optimal decisions by selecting in any given situation S the action A with the highest expected utility

$$MEU_S = \underset{A\in Act_S}{\operatorname{argmax}} \left(p_{A,S}\cdot b_{A,S} - c_{A,S}\right)$$

where Act_S is the set of possible actions in situation S, $p_{A,S}$ is the probability of action A succeeding in S, $b_{A,S}$ is the benefit of A in S in case S succeeds, and $c_{A,S}$ is the cost of attempting A in S (regardless of whether A succeeds).[2] If the agent knows the costs and benefits of each alternative and also the probabilities of each action succeeding in a given situation, it cannot be wrong about which is the most profitable choice. In reality, however, costs and benefits are only approximately known and real-world constraints often can make it difficult or even impossible to estimate accurately the probabilities of action success and failure, which, in turn, eliminates the classical calculation of the expected utility as a viable action selection mechanism. Yet, humans, on the other hand, are subject to the same kinds of real-world constraints, but seem to able to still make good evaluations of situations. Different from classical rational decision processes, human assessments are hypothesized to involve affective states, e.g., so-called “gut feelings” that might encode in some implicit manner how good or bad a situation is for the agent.

To see how affective states and deliberative processes can beneficially be integrated, consider an agent’s “mood state” which is modeled by two state variables, one recording positive affect (Aff_P), the other recording negative affect (Aff_N) (Sloman et al., 2005), both taking values in the interval [0,1]. When a subsystem in the agent architecture performs an activity A successfully, then the level of overall positive affect is increased, and when the subsystem fails, the level of overall negative affect is increased. Specifically, success increases Aff_P by $\Delta Aff_P = (1\text{-}Aff_P)\times inc_A$

[2] Note that we’re modeling action outcomes here as generating benefits only if they succeed, but one could consider a more sophisticated model that includes “unwanted” or “partial” benefits of action failures as potential benefits as well without substantively changing the following discussion.

(failure updates Aff_N analogously), where inc_A is a (possibly learned) value in [0,1] that determines the magnitude of the increase within the available range for the given activity *A*. Both state variables are also subject to regular decay, bringing their activations in the absence of triggering events back to their rest values (i.e., 0): Aff_P is decreased by $\Delta Aff_P = (1-Aff_P) \times dec$ (Scheutz, 2001) (*dec* is the decay value also in [0,1]). Given that affective states encode knowledge of recent success or failures at various architectural activities in a given situation, they can serve as a heuristic that takes past evidence into account in the estimation of action probabilities without the need for an explicit prior distribution of the action probabilities.

We will briefly illustrate this utility of such heuristic estimates in an example taken from (Scheutz & Schermerhorn, 2009). Consider a robot that needs to get to some location and has to decide whether it should ask for directions. The robot does not know that it is in a noisy environment where speech recognition is problematic. All else being equal, the value of inc_A determines how many communication actions *A* the robot will attempt before giving up. With greater inc_A, the value of Aff_N rises faster, leading the robot to reduce its subjective assessment of the expected benefit (i.e., to become “pessimistic” that the benefit will be realized). The robot makes online choices based on the expected utility of a single attempt, using Aff_P and Aff_N to generate an “affective estimate” of the likelihood of success $a = f(Aff_P,Aff_N)$ (*f* here is defined as $f(x,y)=1 + (x^2 - y^2)$ which is then used in the calculation of the expected utility u_A of action *A*: $u_A = a \times p_A \times b_A - c_A$. Note that the effect of positive and negative affect is to modify the benefit the agent expects to receive from attempting the action. When both Aff_P and Aff_N are neutral (i.e., $Aff_P = Aff_N = 0$), the decision is based solely on a comparison of the expected benefit and the action cost. However, given a history of actions, the agent may view the benefit more optimistically (if $Aff_P > Aff_N$) or pessimistically (if $Aff_P < Aff_N$), potentially making decisions that differ from the purely rational choice (overestimating true benefits or costs).

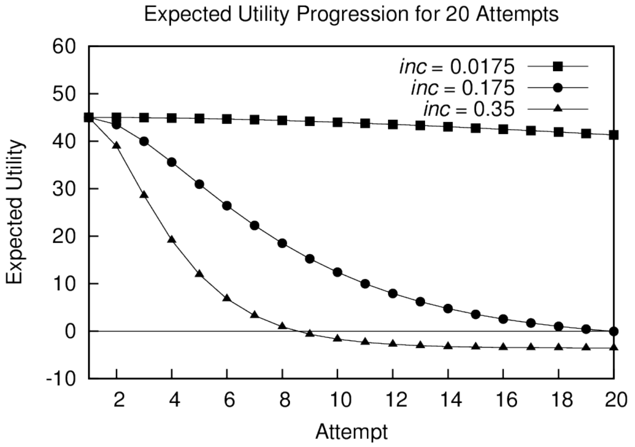

Figure 1: The expected utilities calculated at each communication attempt by the agent for various values of *inc*.

Figure 1 depicts for the communication example the effect of various values of $inc_A = inc$ (for the communication action *A*) on estimates of utility: one that is too optimistic, willing to continue into the foreseeable future; one that is too pessimistic, stopping fairly early; and one that is more reasonable, stopping at about the point where the costs will outweigh the benefits. This suggests that the value of *inc* could be defined as a function of b_A and c_A to improve the likelihood that Aff_N will rise quickly enough to end the series of attempts before costs exceed potential benefits, for example. The agent could employ reinforcement learning to determine the value of inc_A for different actions *A*.

While the value of each state variable is subject to decay, the rate of decay is slow enough that both state variables can serve as *affective memory*, carrying the subjective estimates of the likelihood of success and failure ahead for some period after the events that modified the values of the state variables. In the robot example, after a series of failures leading to the robot deciding not to attempt to ask directions again, the level of Aff_N begins to decay. If, after some period of time, the robot is again faced with the choice of whether to ask for directions, any remaining non-zero value of Aff_N will reduce the likelihood that it will choose to do so. In this way, the robot "remembers" that it has failed recently, and pessimistically "believes" (possibly wrongly) that its chances of failing again are relatively high. Figure 2 shows the expected utility of asking for directions calculated by the robot 100 cycles after a series of failed attempts. The increased "pessimism" leads the evaluation to drop below zero earlier, potentially saving wasted effort on fruitless attempts.

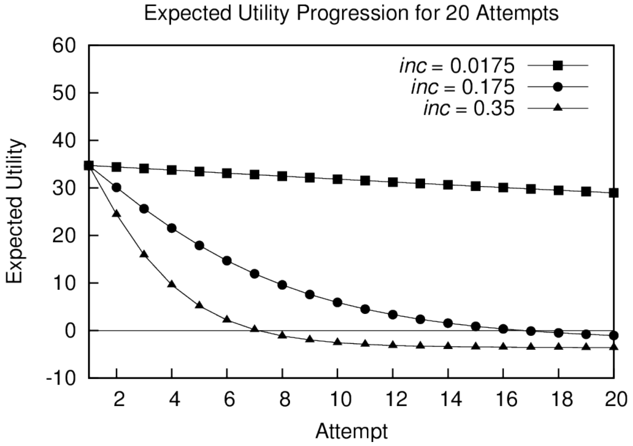

Figure 2: The expected utilities calculated at each attempt by agent for various values of *inc*, *after* an extended series of 20 failures and 100 decay cycles, demonstrating the role of affective states as memory.

It is important to note that affective memory as construed above can conflate different influences and contributing factors. For example, other actions unrelated to speech recognition might have failed as well thus further increasing the robot's negative affect and giving an even more pessimistic outlook than would be warranted by the failed communication attempts alone. Similarly, unrelated actions may have succeeded, increasing positive affect and thus leading to a more positive outlook than warranted by the communication actions. This is consistent with affective evaluations in humans where events unrelated to a particular action can influence overall affective state (e.g., listening to a funny joke can improve, while hearing loud noises can lower, one's overall mood state). And it points to the natural limitations of these simple affect mechanisms that are no substitute for more accurate recordings of $p_{A,S}$ for actions A in situations S; rather, they should be viewed as a "better-than-nothing stand-in" for when the more accurate information is not available, for whatever reason.

Determining the Utility of Affect Mechanisms

One major challenge with affect mechanisms (or any mechanisms, for that matter) in an agent architecture is to assess their utility. With affect the situation is exacerbated by the intrinsic indeterminacy of affect concepts. However, 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 such as fear states, where the functional role is largely well-understood). The strategy we find most useful is based on the view of mental states as intrinsically "architecture-based". As such, mental concepts are analyzed in terms of components and control mechanisms in agent architectures where a mental concept is defined in terms of the minimal set of architectural requirements for the concept to be instantiable in an architecture instance (i.e., running virtual machine, see Sloman, 2002). For this analysis, a framework and language for describing components of agent architectures is needed (Sloman & Scheutz, 2002; Andronache & Scheutz, 2004).

Once affect concepts are analyzed and defined in terms of architectural capacities of agent architectures in such an architecture framework, it is possible to define architectures that implement affective states for a given task and to compare them to agents that implement non-affective architectures for the same task (Scheutz & 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).

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 architecture can have different utilities for different performance measures. Especially in the case of affect, therefore, it is crucial to settle on a notion of utility that reflects our intuitions about why affect should or could be advantageous (e.g., when resources are scarce, fast decisions are required, information is incomplete or unreliable, etc.; Scheutz, 2001). For example, when comparing a “rational agent” playing chess based on minimax search to an affective agent that selects partly suboptimal moves based on emotional states (such as “frustration”, “anger”, “disappointment”, “pleasure”, etc.), the number of games won is not a performance measure that will be particularly useful in determining the utility of affect, because the value of emotional states might not be due to their yielding better absolute performance, but rather better *relative performance*. If it is true (as many psychologists claim) that at least one role of affect is to provide quick and efficient means to reach decisions of importance to an organism that are by and large *good decisions*, 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 (as, indeed, for any architectural construct) the *relative performance-cost trade-off* 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 trade-offs*. To get a handle on the (computational) cost necessary to maintain and run the instantiated architecture, we introduce the notion *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).[3]

Structural costs are those incurred as a result of merely having a certain component or link among components instantiated. They can be thought of as maintenance costs that are associated with any work that needs to be done to keep the architecture instance running. *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 new components, data structures, etc., together with their links, starting, interrupting, or ending processes, etc.). 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 for the complete running architecture instance.

Using the notion of cost induced by an architecture, the notion of *performance-cost-trade-off* is defined as $Perf(P,A,T,E)/Cost(C,A,T,E)$ where $Perf(P,A,T,E)$ is the performance of agent architecture A in task T in environment E under performance measure P and $Cost(C, A,T,E)$ is the cost of operating A during T in E for some cost measure C.[4] Mathematically, performance-cost trade-offs are orders, and can thus form the basis of the comparison of agent architectures: given an order $>_{P,C}$ defined on P and C, an architecture A is said to be *better* than an architecture B with respect to T, E, and P, if $Perf(P,A,T,E)/Cost(C,A,T,E) >_{P,C} Perf(P,A,T,E)/Cost(C,A,T,E)$.

Furthermore, given an architectural parameter λ of an architecture A that can be varied and its set of possible values Λ, we can define an *architecture space* $A_{\lambda,\Lambda}$ and use P to define an order on it. An architecture space $A_{\lambda,\Lambda}$ is said to be *relatively better* than an architecture space $B_{\lambda,\Lambda}$ (with respect to T, E, P, and C), if there exists an architecture A in $A_{\lambda,\Lambda}$ which is better than every

[3] 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).

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

architecture B in $B_{\lambda,\Lambda}$. $A_{\lambda,\Lambda}$ is said to be *absolutely better* than an architecture space $B_{\lambda,\Lambda}$ (with respect to T, E, and P), if $A_{\lambda=c}$ in $A_{\lambda,\Lambda}$ is better than $B_{\lambda=c}$ in $B_{\lambda,\Lambda}$ 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, performance and cost measures). It should be noted that all of the above order notions can be directly extended to sets of tasks, environments, performance and cost measures.

Discussion

We have applied the above methodology of studying affect in artificial agents in several investigations over the last decade with mostly simple simulated agents in the context of biologically plausible survival and procreation tasks that contain foraging and conflict subtasks.[5] Success in those tasks is measured in terms of *the average number of surviving agents of a kind after a large number of agent generations*. The general setup for all studies was a simulated unlimited 2D spatial environment where agents have to forage for food in order to survive and procreate (Scheutz, Schermerhorn, Connaughton, & Dingler, 2006). Initially, specified numbers of agents and food sources are placed in the environment according to a given distribution and then the simulation is run as a *discrete-time simulation* where, at the beginning of each simulation cycle, every agent gets to sense its environment and then decides on an action. All intended actions are then executed in parallel (with the possibility of an action failing if its enabling conditions are not given anymore). Since multiple agents are in the same environment and food is often scarce, conflicts over food can arise, hence agents have to determine whether they want to engage in a conflict over a food item or leave the scene (conflicts over other agents, can also occur, but we are not pursuing this direction here). Simulations are initialized with all initial parameters fixed and then run for a certain number of steps or until some termination criterion is reached (e.g., no remaining live agent). Then different variables in the simulation environment are used for measuring agent performance (e.g., the number of surviving agents, the overall energy stored in agents, etc.). Typical cost measures used in the evaluation were the cost of moving at a given speed or the cost of using certain architectural components. Performance measures are averaged over a set of initial conditions that are taken to be samples from a large space of initial conditions. The averages are then used to perform various statistical analyses (ANOVAs, ANCOVAs, MANOVAs, etc.) in order to determine the dependence of performance on a set of control, bodily, social, and environmental parameters (e.g., Schermerhorn & Scheutz, 2006 and 2007). Instead of reporting the details of the experimental setup together with the specific statistical results here, we will, for space reasons, concentrate on higher-level summaries of our findings, referring the reader to the respective publications for details (see also Scheutz, 2011 for more details on how affective control mechanisms relate to the evolution of communication).

In Scheutz & Sloman (2001), we demonstrated that simple motivational agents (with "hunger-like" and "thirst-like" control mechanisms) are likely to evolve from basic agents under many environmental variations, such as the distribution and influx of energy and water sources in the

[5] We have also started to investigate affective control mechanisms on robots for small architecture spaces, in particular, the types of overall mood states and their interactions with goal management and decision-making as described earlier in the context of simple cooperative human-robot interactions tasks (Schermerhorn & Scheutz, 2009, 2011).

environment as well as the number and distribution of other agents and obstacles. These “hunger-like” and “thirst-like” states were implemented by simple feedback control circuits connected to agent-internal energy-level and water-level sensors and mutation was allowed to operate on the output of these controllers to influence the way in which the control signal was used. In all evolutionary runs, the control output evolved to be used to implement positive time-variant gain values for force vectors pointing to food and water sources. Hence, the control circuit increased the agent’s likelihood of moving towards food or water based on its needs, thus warranting the functional description “hunger-like” and “thirst-like”. Similarly, we demonstrated in Scheutz (2001) that simple “fear-like”, but not “anger-like”, states are likely to evolve, where the labels “fear-like” and “anger-like” were warranted because the agents evolved time-variant gain values for force vectors pointing to other agents and obstacles, thus causing them to move either away from or towards other agents and/or obstacles. We also showed that some of these connections can be learned during an agent’s lifetime using simple associative learning mechanisms (Scheutz, 2000). These results were replicated and extended later in more systematic studies considering larger environmental variations (Scheutz, 2004c, 2004a).

We also investigated the trade-offs between simple reactive agents (with time-invariant gains), simple affective agents (with time-variant gains), and a third class of “deliberative” agents of varying complexity that were able to plan routes through the environment in order to acquire resources more efficiently. In Scheutz, Sloman, & Logan (2000), Scheutz & Logan (2001), and Scheutz & Schermerhorn (2002) we showed that very simple deliberative mechanisms do not pay off in terms of overall performance, especially not if relative performance is considered, i.e., performance where the processing cost of using architectural mechanisms is taken into account – note, again, that relative performance is ultimately what matters for evolutionary considerations because animals will need to spend energy for building, using, and maintaining any additional control circuits in their brains.

In a first attempt to investigate the utility of signaling internal affective states to other agents, we showed that taking other agents’ truthfully-displayed internal fear states into account can lead to significantly better performance in multi-agent foraging tasks where conflicts can arise over resources, as compared to groups that do not indicate their fear levels (Scheutz, 2002). Later, we designed a general game-theoretic framework for conflict resolution in simple agents and showed that there are *fair* conflict resolution strategies (for a particular notion of *fairness*) that lead to Pareto-optimal behavior (Schermerhorn & Scheutz, 2003). Moreover, we showed that there were simple ways of implementing fair strategies based on keeping track of how often an agent won or lost a conflict in the past and making one-shot behavioral decisions about who should get a resource based on this tally, effectively playing an *adaptive rational* strategy. Such adaptive rational agents were superior to all other social and asocial agents in terms of the number of surviving agents after a certain number of generations in the conflict task (Scheutz, 2004b; Scheutz & Schermerhorn, 2004a), including agents that implemented simple dispositional mood-like and motivational states such as overall anxiety or conflict-seeking. Figure 3 shows the performance space comparing the performance of rational adaptive agents to “asocial agents” that have only general dispositional states towards their own kind (*s*-gains) and other kinds (*o*-gains). Depending on the sign of the gain value, these gains implement conflict avoidance or conflict seeking behavior. The particular ranges of *s* and *o* gains depicted here give rise to avoidance of their own kind, but to conflict-seeking with other kinds, thus warranting the labels “fearful of own, aggressive towards other”. As the results show, rational adaptive agents outperform these dispositional asocial agents in the whole parameter space.

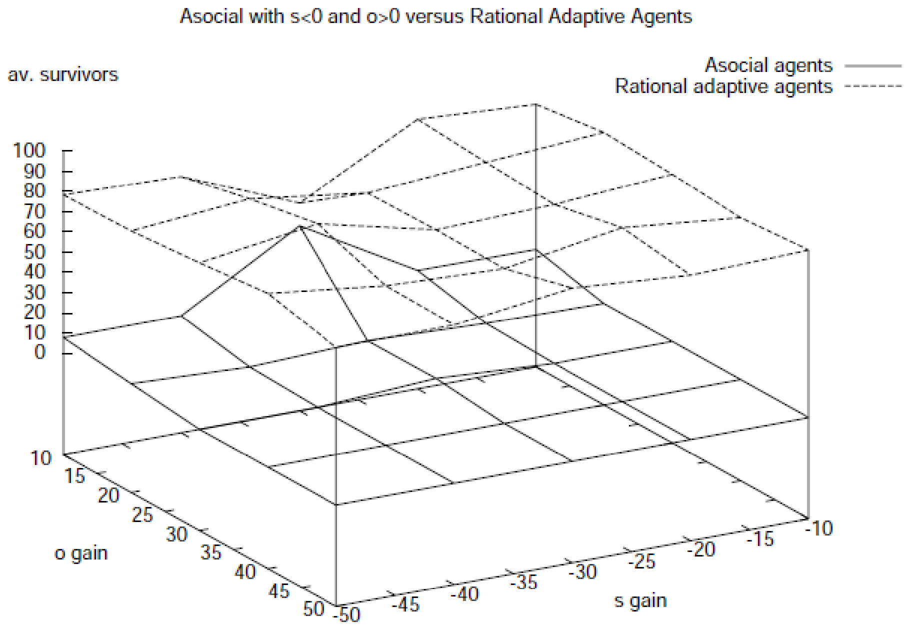

Figure 3: A performance space showing the performance of rational adaptive agents compared to agents with different dispositional states towards their own (*s*-gain) and other (*o*-gain) agent kinds.

When agents are allowed to cheat, i.e., when they can wrongly indicate their action tendencies, then all truth-telling strategies will suffer (Scheutz & Schermerhorn, 2004b), which might be a reason why affective control in nature seems to be largely "hard-coded" to prevent organisms from cheating. We also analyzed the interactions between simple non-social affective control and social control through affective displays and conflict resolution strategies in order to determine the trade-offs between individual and social strategies and found that agents could make up for suboptimal strategies in the conflict task using specific gain values (behavioral propensities) in their foraging control that allowed them to avoid conflicts more frequently (Scheutz, 2006), thus providing evidence for the utility of simple (non-social) affective control (in the foraging task) in the light of conflicts, also possibly providing a way for agents to cope with cheaters in conflict tasks.

It is important to note that most of the above results did not attempt a detailed break-down of the costs induced by the architectures, but rather used a notion of (compounded) overall cost. It would thus be useful to repeat the previous experiments with a much more detailed analysis in terms of structural, process, and action costs to obtain 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 trade-off" between structural and processing costs (e.g., the difference between an affective 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).

Conclusions

In this paper, we argued for a systematic architecture-based research program to investigate the utility of affect mechanisms for simulated and robotic agents. We gave examples of functional and architecture-based characterizations of simple and complex motivational and emotional states and also discussed how affective states could usefully interact with rational processes. We then introduced a framework for performing systematic experiments with artificial agents based on the notion of cost induced by an architecture. The latter can be used to determine relative performance-cost trade-offs among different architectural mechanisms (affective and non-affective) and provides an answer to the previously-raised question about when to use affective agents: we should consider affect mechanisms over other non-affective mechanisms for a given problem (even the provably best ones) when the relative performance of the affective solution is better for the given tasks in the given environments. To be able to give this kind of answer for a set of agents, is the aim of the proposed program of mapping out the performance spaces of affective agents for a large number of tasks and environments. And even though it is much easier to perform these types of evaluations with virtual agents in simulated environments, there is nothing in principle that prevents us from performing the same kinds of experiments with robots in real environments (in many cases, of course, the complexity and duration of such investigations will be prohibitive, in which case accurate simulations might be able to serve as a reasonable, acceptable substitute). Ultimately, this kind of systematic evaluation of affect mechanisms in different agent architectures across different tasks and environments is necessary for us to be able to make *general statements about the utility of affect for artificial agents*.

Acknowledgments

The author would like to thank Paul Schermerhorn for helpful suggestions to improve the narrative and Aaron Sloman for his influential work on architecture-based affect concepts.

References

Allen, S. (2001). *Concern processing in autonomous agents*. Unpublished doctoral dissertation, School of Computer Science, The University of Birmingham.

Andronache, V., & Scheutz, M. (2004). Integrating theory and practice: The agent architecture framework APOC and its development environment ADE. In *Proceedings of aamas 2004* (pp. 1014–1021). ACM Press.

Arkin, R., Fujita, M., Takagi, T., & Hasegawa, R. (2003, March). An ethological and emotional basis for human-robot interaction. *Robotics and Autonomous Systems*, *42*, 3-4.

Beaudoin, L., & Sloman, A. (1993). A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, & A. Ramsay (Eds.), *Prospects for artificial intelligence* (pp. 229–238). Amsterdam: IOS Press.

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

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

Bless, H., Schwarz, N., & Wieland, R. (1996). Mood and the impact of category membership and individuating information. *European Journal of Social Psychology*, *26*, 935-959.

Bower, G. H., & Cohen, P. R. (1982). Emotional influences in memory and thinking: Data and theory. In M. Clark & S. T. Fiske (Eds.), *In affect and cognition: The seventh annual carnegie symposium on cognition* (pp. 291–331). N.J: Lawrence Erlbaum Associates.

Breazeal, C. L. (2002). *Designing sociable robots*. MIT Press.

Cãnamero, D. (1997). Modeling motivations and emotions as a basis for intelligent behavior. In L. Johnson (Ed.), *Proceedings of the first international symposium on autonomous agents (agents'97)* (pp. 148–155). New York, NY: ACM Press.

Clore, G., Gasper, K., & Conway, H. (2001). Affect as information. In J. Forgas (Ed.), *Handbook of affect and social cognition* (p. 121-144). Mahwah, NJ: Erlbaum.

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (2^{nd} ed., pp. 91–115). NY: Guilford.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Gosset/Putnam Press.

Davidson, R. J., Scherer, K. R., & Goldsmith, H. H. (Eds.). (2003). *Handbook of affective sciences*. New York: Oxford University Press.

Derryberry, D., & Tucker, D. (1994). Motivating the focus of attention. In P. Neidenthal & S. Kitayama (Eds.), *The heart's eye: Emotional influence in perception and attention* (pp. 67–96). San Diego, CA: Academic Press.

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

Eberbach, E. (2001). Evolutionary computation as a multi-agent search: A $-calculus perspective for its completeness an d optimality. In *Proceedings of congress on evolutionary computation cec'2001* (p. 823-830). Seoul, Korea.

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

Ekman, P., & Davidson, R. J. (Eds.). (1994). *The nature of emotion: Fundamental questions*. New York: Oxford University Press.

Eliott, C. (1992). *The affective reasoner: A process model of emotions in a multi-agent system*. Unpublished doctoral dissertation, Institute for the Learning Sciences,Northwestern University.

El-Nasr, M., Yen, J., & Ioerger, T. (2000). Flame – fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, *3*(3), 219-257.

Franklin, S., Kelemen, A., & McCauley, L. (1998). Ida: a cognitive agent architecture. In *Ieee conference on systems, man and cybernetics* (pp. 2646–2651).

Fredrickson, B. (1998). What good are positive emotions? *Review of General Psychology, 2*, 300–319.

Frijda, N. H. (1986). *The emotions. Studies in emotion and social interaction*. Cambridge: Cambridge University Press.

Frijda, N. H. (2000). The psychologists' point of view. In (Lewis & Haviland-Jones, 2000) (pp. 59–74).

Gadanho, S. C. (2003). Learning behavior-selection by emotions and cognition in a multi-goal robot task. *Journal of Machine Learning Research*, *4*(Jul), 385-412.

Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research*, *5*(4), 269–306.

Gray, J. (1990). Brain systems that mediate both emotions and cognitions. In J. Gray (Ed.), *Psychobiological aspects of relationships between emotion and cognition.* Hillsdale, NJ: Lawrence Erlbaum.

Hamm, A. O., Schupp, H. T., & Weike, A. I. (2003). Motivational organization of emotions: Autonomic changes, cortical responses, and reflex modulation. In (Davidson et al., 2003) (pp. 187–211).

Hayes-Roth, B., Ball, G., Picard, R. W., Lisetti, C., & Stern, A. (1998). Panel on affect and emotion in the user interface. In *International conference on intelligent user interfaces* (pp. 91–94). New York: ACM Press.

Humphrey, N. (1992). *A history of the mind*. Copernicus Books.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review*, *100*(1), 68–90.

Kahneman, D., Wakker, P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, *112*, 375-405.

LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.

LeDoux, J. E., & Fellous, J. (1995). The handbook of brain theory and neural networks. In M. A. Arbib (Ed.), *Emotion and computational neuroscience* (p. 356-360). Cambridge, MA: MIT Press.

Lewis, M., & Haviland-Jones, J. M. (Eds.). (2000). *Handbook of emotions* (2nd ed.). New York: The Guilford Press.

Lorenz, K., & Leyhausen, P. (1973). *Motivation and animal behavior: An ethological view*. New York: Van Nostrand Co.

Lorenz, K. Z. (1981). *The foundations of ethology*. Springer-Verlag, New York.

McFarland, D. (1981). *The oxford companion to animal behavior*. Oxford: Oxford University Press.

Michaud, F., & Audet, J. (2001). Using motives and artificial emotion for long-term activity of an autonomous robot. In *Proceedings of the 5th autonomous agents conference* (pp. 188–189). Montreal, Quebec: ACM Press.

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

Mueller, E. T. (1998). *Natural language processing with thoughttreasure*. New York: Signiform.

Murphy, R. R., Lisetti, C., Tardif, R., Irish, L., & Gage, A. (2002). Emotion-based control of cooperating heterogeneous mobile robots. *IEEE Transactions on Robotics and Automation*, *18*(5), 744-757.

Nourbakhsh, I., Bobenage, J., Grange, S., Lutz, R., Meyer, R., & Soto, A. (1999). An affective mobile educator with a full-time job. *Artificial Intelligence*, *114*((1-2)), 95-124.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of the emotions*. New York: Cambridge University Press.

Panksepp, J. (2000). Emotions as natural kinds within the mammalian brain. In (Lewis & Haviland-Jones, 2000) (pp. 137–156).

Picard, R. (1997). *Affective computing*. Cambridge, Mass, London, England: MIT Press.

Scherer, K., Schorr, A., & Johnstone, T. (2001). *Appraisal theories of emotions: Theories, methods, research*. New York: Oxford University Press.

Schermerhorn, P., & Scheutz, M. (2003). Implicit cooperation in conflict resolution for simple agents. In *Agent 2003.* Chicago, IL: University of Chicago.

Schermerhorn, P., & Scheutz, M. (2006, May). Social coordination without communication in multi-agent territory exploration tasks. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-06)* (pp. 654–661). Hakodate, Japan.

Schermerhorn, P., & Scheutz, M. (2007, April). Social, physical, and computational tradeoffs in collaborative multi-agent territory exploration tasks. In *Proceedings of the First IEEE Symposium on Artificial Life* (pp. 295–302).

Schermerhorn, P., & Scheutz, M. (2009, November). Dynamic robot autonomy: Investigating the effects of robot decision-making in a human-robot team task. In *Proceedings of the 2009 International Conference on Multimodal Interfaces.* Cambridge, MA.

Schermerhorn, P., & Scheutz, M. (2011). Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task. In *The Fourth International Conference on Advances in Computer-Human Interactions* (pp. 236–241).

Scheutz, M. (2000). Surviving in a hostile multiagent environment: How simple affective states can aid in the competition for resources. In H. J. Hamilton (Ed.), *Advances in Artificial Intelligence, 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000, Montréal* (Vol. 1822, pp. 389–399). Springer.

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

Scheutz, M. (2002). The evolution of affective states and social control. In C. K. Hemelrijk (Ed.), *Proceedings of International Workshop on Self-organisation and Evolution of Social Behaviour* (pp. 358–367). Monte Verità, Switzerland.

Scheutz, M. (2004a). An artificial life approach to the study of basic emotions. In *Proceedings of Cognitive Science 2004.*

Scheutz, M. (2004b). On the utility of adaptation vs. signalling action tendencies in the competition for resources. In *Proceedings of aamas 2004* (pp. 1378–1379). ACM Press.

Scheutz, M. (2004c). Useful roles of emotions in artificial agents: A case study from artificial life. In *Proceedings of AAAI 2004* (pp. 31–40). AAAI Press.

Scheutz, M. (2006, June). Cross-level interactions between conflict resolution and survival games. In *Proceedings of Artificial Life X* (pp. 459–465).

Scheutz, M. (2011). Evolution of affect and communication. In D. Gökcay & G. Yildirim (Eds.), *Affective computing and interaction: Psychological, cognitive and neuroscientific perspectives* (pp. 75–93).

Scheutz, M., & Logan, B. (2001). Affective versus deliberative agent control. In S. Colton (Ed.), *Proceedings of the AISB'01 Symposium on Emotion, Cognition and Affective Computing* (pp. 1–10). York: Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Scheutz, M., & Schermerhorn, P. (2002). Steps towards a theory of possible trajectories from reactive to deliberative control systems. In R. Standish (Ed.), *Proceedings of the 8th Conference of Artificial Life* (pp. 283–292). MIT Press.

Scheutz, M., & Schermerhorn, P. (2004a). The more radical, the better: Investigating the utility of aggression in the competition among different agent kinds. In *Proceedings of SAB 2004* (pp. 445–454). MIT Press.

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

Scheutz, M., & Schermerhorn, P. (2009). Affective goal and task selection for social robots. In J. Vallverdú & D. Casacuberta (Eds.), *The handbook of Research on Synthetic Emotions and Sociable Robotics.* IGI Global.

Scheutz, M., Schermerhorn, P., Connaughton, R., & Dingler, A. (2006, June). SWAGES–an extendable parallel grid experimentation system for large-scale agent-based alife simulations. In *Proceedings of artificial life x* (pp. 412–418).

Scheutz, M., Schermerhorn, P., Kramer, J., & Middendorff, C. (2006). The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM International Conference on Human-Robot Interaction* (pp. 226–233).

Scheutz, M., & Sloman, A. (2001). Affect and agent control: Experiments with simple affective states. In N. Zhong, J. Liu, S. Ohsuga, & J. Bradshaw (Eds.), *Intelligent Agent Technology: Research and Development* (pp. 200–209). New Jersey: World Scientific Publisher.

Scheutz, M., Sloman, A., & Logan, B. (2000). Emotional states and realistic agent behaviour. In P. Geril (Ed.), *Proceedings of Gameon 2000, Imperial College London* (pp. 81–88). Delft: Society for Computer Simulation.

Schwarz, N. (1990). Feelings as information: Informational and motivational functions of affective states. In E. Higgins & R. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of Social Behavior* (Vol. 2, p. 121-144). New York: Guilford Press.

Searle, J. R. (1992). *The rediscovery of the mind*.

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

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies, 10*(4-5).

Sloman, A., Chrisley, R., & Scheutz, M. (2005). The architectural basis of affective states and processes. In J. Fellous & M. Arbib (Eds.), *Who needs emotions? The brain meets the machine.* New York: Oxford University Press.

Sloman, A., & Scheutz, M. (2002). A framework for comparing agent architectures. In *UK Workshop on Computational Intelligence* (pp. 169–176).

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

Wehrle, T., & Scherer, K. (2001). Towards computational modeling of appraisal theories. In (Scherer et al., 2001) (p. 350-365).

Wright, I., Sloman, A., & Beaudoin, L. (1996). Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, *3*(2), 101–126. (Repr. In R.L.Chrisley (Ed.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, Vol IV, Routledge, London, 2000)