First Steps toward Natural Human-Like HRI

M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson
Artificial Intelligence and Robotics Laboratory
Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556, USA
{mscheutz,pscherm1,jkramer3,danderso}@cse.nd.edu

Abstract

Natural human-like human-robot interaction (NHL-HRI) requires the robot to be skilled both at recognizing and producing many subtle human behaviors, often taken for granted by humans. We suggest a rough division of these requirements for NHL-HRI into three classes of properties: (1) *social behaviors*, (2) *goal-oriented cognition*, and (3) *robust intelligence*, and present the novel DIARC architecture for complex affective robots for human-robot interaction, which aims to meet some of those requirements. We briefly describe the functional properties of DIARC and its implementation in our ADE system. Then we report results from human subject evaluations in the laboratory as well as our experiences with the robot running ADE at the 2005 AAAI Robot Competition in the *Open Interaction Event* and *Robot Exhibition*.

1 Introduction

We take the ultimate goal of human robot interaction (HRI) to be the achievement of *natural* and *human-like* (NHL) robot behavior as it relates to human contact. By this, we mean that our intention is to establish a robotic architecture for HRI such that any restrictions on possible interactions are due to *human capacities* (i.e., the limitations of human perceptual, motor, or cognitive system), and not the *a priori* functionality of the robot. For example, an interaction that would require humans to speak at ten times the normal speech rate would be excluded, as would interactions that required humans to hear ultrasound or communicate in some non-human language. However, we do want to include interactions that can occur in any typical human setting, such as the ability to use language freely in any way, shape, or form; to make reference to personal, social, and cultural knowledge; or to involve all aspects of human perception and motor capabilities.

Clearly, NHL-HRI is not an achievable goal for any robotic system in the foreseeable future; in fact, we are not even close. Yet, we believe that is not too early to start a discussion of possible requirements for NHL-HRI, given current achievements in HRI and knowledge of robotic architectures. Specifically, we believe that it will be critical for future robotic efforts in HRI to investigate architectural structures, principles, and concepts that necessarily (or even potentially) have a role in robots capable of NHL-HRI.

In this paper, then, we will start with a modest reflection on three classes of properties that we deem crucial to successful natural HRI, with an eye towards NHL-HRI. We then present a brief overview of our own attempts at defining a Distributed Integrated Affect, Reflection, and Cognition architecture (DIARC), as a first step towards architectures for NHL-HRI. After discussing the functional organization of DIARC and the role of affect in the integration of its various subsystems, we briefly describe some of the implemented components, and compare results from testing DIARC in human subject experiments in the laboratory with our experiences of the robot's interactions with people in the real-world context of the AAAI 2005 robot competition, specifically the *Open Interaction Event* and *Robot Exhibition*. Finally, we provide a summary of the current state of DIARC and conclude with some thoughts on the role of affect in NHL-HRI-capable robots.

2 The DIARC Architecture

Natural human-like human-robot interaction places many complex demands on a robotic architecture. As a first rough cut, we can divide them into three main categories: (1) *social behaviors*, (2) *goal-oriented cognition*, and (3) *robust intelligence*. The first category includes all aspects of human communicative acts, the second is concerned with all forms human cognition and teleological behavior, and the third centers around various mechanisms that ensure the reliable, long-term, fault-tolerant autonomy and survival of the robot. We will now briefly expand on each of these three categories.

First and foremost, it is immediately clear that robots must be capable of natural language processing if humans are to be free to use (spoken) language whenever and in whatever form they please. In addition to speech recognition and production, natural language interactions will require methods for semantic processing of language structures and natural language understanding. Moreover, knowledge of dialog structures, dialog progression, and teleological discourse is required for the robot to be able to engage in natural communicative interaction patterns (e.g., Grosz & Sidner, 1990). This also includes the recognition of human affect and the appropriate expression of affect on the part of the robot (e.g., in response to recognized affect), as well as mechanisms for recognizing and producing other non-verbal cues (e.g., gestures, head movements, gaze, etc.) that accompany human discourse and social interactions.

Second, genuine natural interactions require the robot to communicate, instill, and prompt the *ascription of intentionality* (i.e., the human ability to treat systems as if they had their own intentions). Human interlocutors will automatically watch for behaviors that convey intent (e.g., as established by non-verbal cues) and assume that the robot has the ability to recognize and utilize such behaviors in others. As part of being able to present itself *consistently and over extended periods of time* as a purpose-driven entity, the robot will require *genuine purposes*, represented as goals and implemented in internal goal and task management mechanisms.[1] Their absence will have disruptive effects on the natural flow of conversation and, eventually, the overall interaction. Ultimately, the robot has to behave in a way that supports a consistent human "theory of robot minds," which is the human ascription of human-like beliefs, intentions, and desires that make the robot predictable to humans.

Third, the architecture must include mechanisms to recover both from failures within the system (e.g., acoustic, syntactic, semantic misunderstandings, dialog failures, etc.) as well as failures of the system itself (e.g., crashes of components, internal timing problems, faulty hardware, etc.).

2.1 The Utility of Affect for Controlling Information Flow in Agent Architectures

Affective robotic architectures can directly address many demands for HRI, in addition to providing many benefits for other robot tasks. Affect plays a critical role for humans in social interactions; for instance, speech recognition, even when perfect, makes up only one part of the meaning of an utterance. Expressions of affect (e.g., via gestures, facial expressions, or prosodic characteristics, see Ekman, 1993) augment or modify the explicit semantic content of statements (e.g., a sarcastic tone might indicate that the speaker's belief is the opposite of the spoken content, or emphatic gestures might indicate the strength of the speaker's belief). To create accurate representations of others' mental states based on conversation, it is necessary to "pick up on" these cues. Moreover, humans expect and look for these cues when processing the robot's speech output, so the inclusion of affect expression capabilities should enhance the degree to which humans feel *they* can accurately assess *the robot's* belief states (the utility of affect expression will be demonstrated in Section 3.1).

Affect can also be an effective way for higher-level deliberative mechanisms in an agent architecture to connect to and utilize motivational mechanisms of lower-level non-deliberative components. Specifically, deliberative mechanisms can alter the states of these components (e.g., by injecting new "force" into "affective circuits" or by suppressing output of those circuits) to create and modify existing goals, directly influence the control of action (e.g., by changing the preferences in the agent's action selection mechanism), and drive learning based on internally generated valuations and value signals (Scheutz, 2000). Thus, affect may serve the purpose of integration and management of multiple processes required for the effective functioning of an autonomous system (Ortony, Norman, & Revelle, 2005); affect allows for motivational signals originating not from changes in the external environment detected via sensors, but from components within the architecture itself (e.g., from deliberative subsystems). Such signals can then influence various other parts of the architecture and modify goal management, action selection, and learning (e.g., see Scheutz, 2004, where we isolated 12 functional roles of emotions in an agent architecture).

Finally, whereas a complex cognitive evaluation of a situation may provide a more accurate assessment (and likely a better subsequent action selection) than that provided by an affective evaluation, it may prove too costly or time-consuming to be practical for real-time use. For fast, low-cost approximations, humans seem to use *affective memory* (Bless, Schwarz, & Wieland, 1996), which encodes implicit knowledge about the likelihood of occurrence of a positive or negative future event (Clore, Gasper, & Conway, 2001). Robotic agents can use such affective states as subjective probabilities in *affective evaluations* of potential actions. This can be useful in quickly determining an appropriate reaction to catastrophic failures in the system (e.g., the failure of the vision subsystem).

[1] Note that the emphasis here is on both "consistency" and "extended time period" as humans can sometimes be tricked into believing that something has a purpose because it *seems* to exhibit purposeful behavior for short periods of time. However, the deception will typically not last for long (e.g., see the repeatedly failed attempts at convincing humans that a computer is a human in the Loebner prize competition).

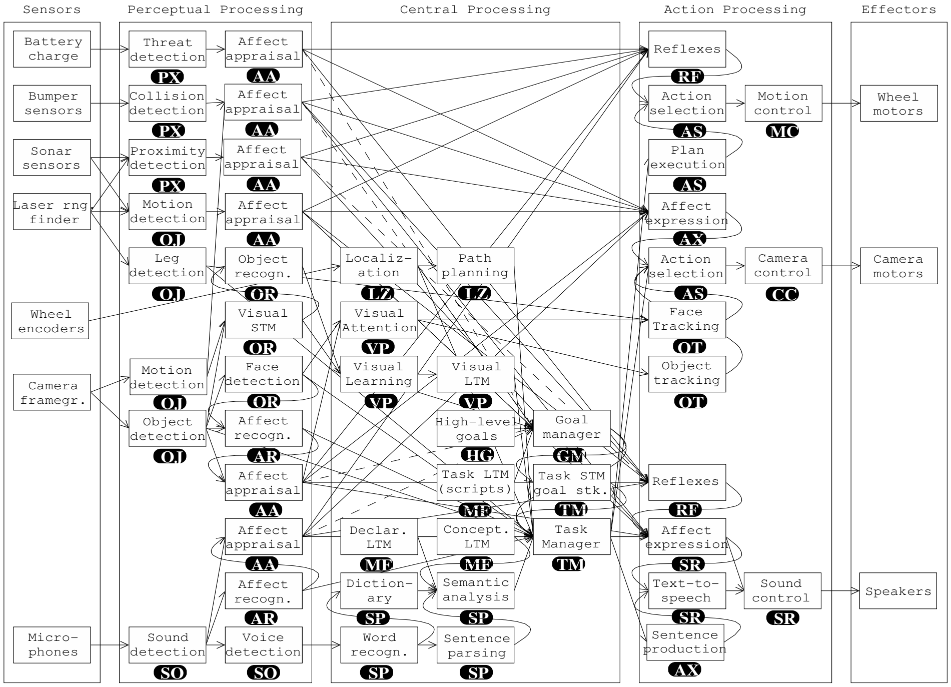

Figure 1: A high-level view of the functional components in the proposed DIARC architecture for complex human-like robots. Boxes depict concurrently running components of varying complexity. Solid arrows depict information flow and dashed arrows depict control flow through the architecture (the latter via different affective processes). Only links pertaining to affect processing are shown. Labels (black ovals) are included to relate architectural components to their counterparts in the implementation diagram shown in Figure 2.

2.2 A Brief Overview of DIARC

Figure 1 depicts a partial view of the functional organization of the proposed affective architecture for complex robots: DIARC, the Distributed Integrated Affect, Reflection, Cognition architecture. Sensors, effectors, and perceptual, central, and action processing components are separated into columns. All boxes depict autonomous computational units that can operate in parallel and communicate via several types of communication links. Names of components denote their functional roles in the overall system.

At a high level, an agent that implements DIARC operates as follows: sensory information is gathered via the various sensors and passed on the the appropriate perceptual processing components. In perceptual processing, raw sensory input is (potentially) parsed into meaningful perceptual data. Much of these data are accessed by elements of the central processing group for goal and task management. In addition, perceptual data are routed to affective appraisal modules, which make fast evaluations of the potential (positive or negative) effects they may imply for the robot. This may lead, in turn, to "reflexive" reactions that bypass the entire central processing step (e.g., causing the robot to stop immediately in response to a perceived threat, rather than considering the costs and benefits of stopping, comparing those to the costs and benefits of alternative actions, and selecting the best action overall). The central processing step performs action selection based on the perceptual input, concept memory, cost-benefit analysis, etc. At the level of action processing, directives from the perceptual and central processing stages generate commands for the effectors, with conflicts generally resolved in favor of commands from perceptual processing (e.g., the reflexive STOP command overrides movement commands generated by plan execution). This is accomplished by priority-based action selection mechanisms (e.g., Scheutz & Andronache, 2004).

2.3 The Vision System

The *vision system* has two components: a *social* component for detecting and tracking faces and facial features and a *recognition* component for object recognition, learning, and memorization. The social subsystem, for example, provides information about detected faces using histogram methods from Yang, Kriegman, and Ahuja (2002) as well as skin color, color of clothes, and the camera pan and tilt angles, which are used in conjunction with information about the distance of an object (as reported by sonar and laser sensors) to identify and track people in a room (Scheutz, McRaven, & Cserey, 2004) and determine some of their salient features such as height, based on distance and camera tilt angle (Byers, Dixon, Goodier, Grimm, & Smart, 2003) for future identification. Feature extraction of the eyes and mouth is performed using a swarm-based exploration of a parameter space to find parameters for a Canny edge detector that produce the best features within a face region given by the OpenCV Haar cascade (Middendorff & Scheutz, 2006). Motion of the tracked features gives insight into emotional changes of the subject, as seen in Ekman and Friesen (1977). For example, a sudden raise in the eyebrows can signal happiness or surprise, whereas downward movement can indicate frustration or confusion. The general shapes and positions of features are maintained in a hash table keyed by face and leg positions, allowing emotional states to be tracked simultaneously for several people. The emotion recognition component currently only classifies faces as "happy", "sad" or "neutral".

Object recognition was implemented using *scale-invariant feature detection* (SIFT) points for the extraction of distinctive features from images (Lowe, 2004). SIFT keypoints are invariant with respect to image scale, rotation, change in 3D viewpoint and change in illumination, and can be used to perform a variety of tasks including reliable object recognition, even in situations where the scene is cluttered or the object is partially occluded.

Object recognition is a two-phase process for unknown objects, composed of *learning* an object and subsequent *memory recall*. When learning an object, the SIFT keypoints of an image are translated to an ASCII representation[2] and a distinct identification token is stored in a database. During recall, the system generates an ASCII representation of the keypoints in the current scene. This representation is compared to entries in the database using a nearest-neighbor algorithm, allowing the system to answer queries such as "Is object X in the scene?" and "Which objects in the database are in the scene?" Once objects have been positively identified, it is possible to perform additional queries concerning their spatial relationships (e.g., "Is object X to the left of object Y?" etc.).

SIFT-based object recognition performs most effectively when attempting to identify rigid objects with a distinctive pattern. It reliably differentiates between the covers of various textbooks, assorted computer components and the boxes of various household products. However, fewer keypoints are detected on curved surfaces, with a correspondingly lower recognition rate for objects such as balls or aluminum cans, made worse if the objects lack distinctive markings and surface features.

2.4 Natural Language Processing

The *natural language processing subsystem* integrates and extends various extant components. Speech recognition is performed using SONIC (Pellom & Hacioglu, 2003) and Sphinx (The Sphinx Group at Carnegie Mellon University, 2004); parsing is handled by the link parser (Sleator & Temperley, 1993). Verbnet (Kipper, Dang, & Palmer, 2005) and Framenet (Fillmore, Baker, & Sato, 2002) are used for semantic processing, in conjunction with a modified version of Thought Treasure for natural language understanding and speech production (Mueller, 1998). Speech synthesis is handled by the University of Edinburgh's *Festival* system (Festival, 2004), augmented by an emotional output filter (Burkhart, 2005). In addition, our own components are employed for affect expression in spoken language (e.g., "angry", "frightened", "happy", "sad", and their gradations, such as "halfangry", which indicates a somewhat elevated state of anger).

2.5 Action Interpretation and Selection

The *action control subsystem* (Scheutz, Schermerhorn, Kramer, & Middendorff, 2006) is based on a novel *affective action interpreter*, which interprets scripts (Schank & Abelson, 1977) stored in long-term memory. Scripts encode the robot's procedural knowledge of certain conversation "templates" as well as complex action sequences for task performance. These scripts can be combined in hierarchical and recursive ways, yielding complex behaviors from basic behavioral primitives, which are grounded in basic *skills* (Ichise, Shapiro, & Langley, 2002). Moreover, several spatial maps are used for the representation of locations of the robot, people, and other salient objects in the environment, as well as for path planning and high-level navigation. As a trivial example, the following script (activated from a higher-level script as `notify-beverage-done(self, Jim, coffee, desk3)`) instructs the robot to move to location `desk3` (retrieved from a map in long-term memory) shift its focus of attention to the human there, and tell the human that his coffee is ready:

```
====notify-beverage-ready//serve.V
role01-of=robot|
role02-of=human|
role03-of=beverage|
role04-of=loc|
timeout-of=600sec|
event01-of=[move-to robot loc]|
event02-of=[shiftFOA robot human]|
event03-of=[say-to robot human
            [human your beverage is ready.]]|
```

The action interpreter substitutes the arguments passed for each of the roles in the script and begins executing the

[2] The ASCII representation is generated with the binary program distributed on Lowe's website http://www.cs.ubc.ca/∼lowe/keypoints/.

first event. A behavioral primitive like `move-to(robot, loc)` then has a particular meaning to the robotic system. In this case, the action interpreter passes the action on to the navigation system, which interprets it as a command to move the robot to the coordinates (x,y) of `loc` (represented in a discrete, topological map). The high-level navigation system generates a plan which translates the action into commands for the low-level navigation system, eventually causing the robot to move in a particular direction, if possible (e.g., it will not move there if obstacles block the location, although it will attempt to move around obstacles that obstruct the path to the final location).

In addition to action primitives or references to other scripts, scripts also support conditional execution whereby the next step is determined by the outcome of the present event. In this way, failure recovery actions can be encoded directly in the scripts. For example, suppose there were no `human` at `desk3` in the above example. In that case, the `shiftFOA(human)` goal cannot be achieved, so a more appropriate action would be to end `notify-beverage-ready` in its failure state (which can in turn be detected and addressed appropriately by the calling script), rather than delivering the message to nobody and indicating successful completion.

A failure such as the one described above causes an adjustment to the robot's affective state, which subsequently allows for an additional context-based adaptation of goals, preferences, attitudes, and, ultimately, behavior. Affective states do not explicitly influence action selection (i.e., there is no branching in scripts based on affective states). Instead, affect influences the priorities assigned to the goals currently held by agent. For each goal currently held by the agent, there is an associated script interpreter that manages the execution of script events to achieve that goal. These script interpreters execute concurrently, so multiple goals may be advanced at the same time. However, when conflicts arise (e.g., when two scripts require the same physical resource, such as motor control), they are resolved in favor of the goal with the highest priority. A goal's priority is based on its *importance* (i.e., benefits minus cost scaled by affective evaluations, see Scheutz et al., 2006) and its *urgency*, which reflects the likelihood that there will be sufficient time remaining to complete the script. By mediating the importance component of goal priority, affect can effectively alter the robot's perception of a goal's utility. Positive affect leads to more "optimistic" assessments of utility (and, hence, higher priority), whereas negative affect leads to more "pessimistic" assessments of utility.[3]

2.6 System Infrastructure

There are certain aspects of a complex robot that are critical for long-term, safe, and flexible operation. The computational demands of various sub-systems require distributing the architecture across hosts, allowing concurrent operation while retaining system reactivity. Furthermore, the importance of monitoring and maintaining system integrity and health (including error detection, system reconfiguration, and failure recovery) cannot be overstated. DIARC is implemented within ADE, the *Architecture Development Environment* (Andronache & Scheutz, 2006; Scheutz, 2006), which provides a multi-agent based *infrastructure*. ADE is not an architecture itself, but a *framework* for developing, debugging, and deploying complex agent architectures based on the APOC universal agent architecture formalism (Scheutz & Andronache, 2003; Andronache & Scheutz, 2004); ADE incorporates various tools and mechanisms that promote reliable and flexible system operation (Kramer & Scheutz, 2006b, 2006a).

The basic component in ADE is the ADESERVER, which is comprised of one or more computational processes that serve requests. Accessing the services provided by an ADESERVER is accomplished by obtaining a *reference* to the (possibly remote) server, forming a *local representation* that is referred to as an ADECLIENT. The ADEREGISTRY, a special type of ADESERVER, mediates connections among servers and the processes that use their services. In particular, it organizes, tracks, and controls access to servers that register with it, acting in a role similar to a *white-pages service* found in multi-agent systems. The ADEREGISTRY provides the backbone of an ADE system; all components must register to become part of the architecture. A set of components may contain multiple registries that mutually register with one another to provide both redundancy and the means to maintain distributed knowledge about the system.

All connected components of an implemented architecture (that is, ADESERVERS, ADECLIENTS, and ADEREGISTRYS) maintain a communication link during system operation. At a bare minimum, this consists of a periodic *heartbeat* signal indicating that a component is still functioning. A server sends a heartbeat to the registry with which it is registered, while a client sends a heartbeat to its originating server. The component receiving the heartbeat periodically checks for a heartbeat signal; if none arrives, the sending component receives an error, while the receiving component times out. An ADEREGISTRY uses this information to determine the status of servers, which in turn determines their accessibility. An

[3] The current implementation of the action interpreter is still somewhat impoverished, as variables for other scripts have not been implemented yet. For example, it is not possible to add "variable actions" to scripts such as "pick any script that satisfies preconditions X_i and execute it", which would cause the action interpreter to search through its scripts and match them against the preconditions X_i. Also, the current implementation only supports detection of failures, but not "recursive" attempts to recover from them – "recursive", for recovery actions might themselves fail and might thus lead to recovery from recovery, etc.

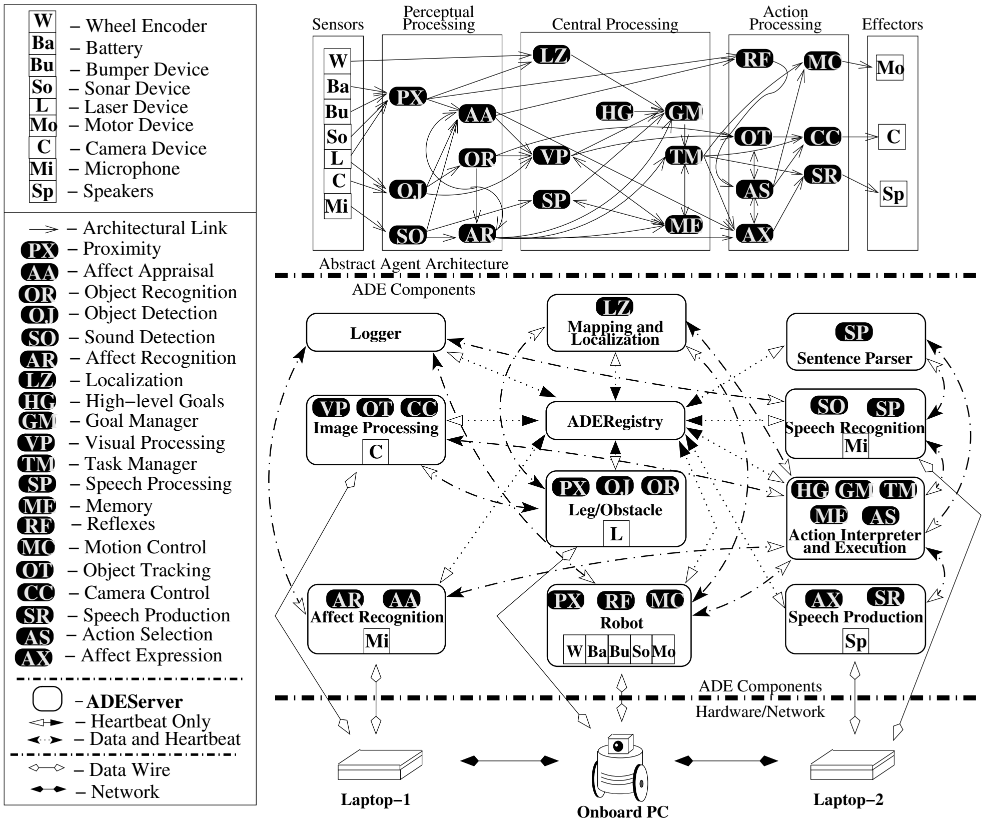

Figure 2: Three representations of the proposed DIARC architecture for complex human-like robots. The bottom level depicts the system (or hardware), the middle level depicts the multi-agent system (or ADE components), and the top level depicts a simplified view of the DIARC agent architecture shown in Figure 1. Note that the ADEREGISTRY and "Logger" components are part of the infrastructure, not part of DIARC itself.

ADESERVER uses heartbeat signals to determine the status of its clients, which can then determine if the server's services remain available.

Figure 2 shows a "3-level" view of the architecture; the top depicts a partial view of the abstract DIARC architecture, while the middle and bottom depict its breakdown into components in ADE and the hardware on which it executes, respectively. The robot platform used is an ActivMedia Peoplebot with an on-board computer and two additional laptops (all running Linux with a 2.6.x kernel). Available hardware includes a pan-tilt-zoom camera, a SICK laser range finder, three sonar rings, two microphones, two speakers, a local Ethernet network, and one wireless link to the "outside world."

3 DIARC Applications

Over the last two years, DIARC has been used in various applications, from informal laboratory evaluations, to human subject experiments, to public demonstrations and robot competitions to evaluate the functionality of different subsystems of DIARC as well as the viability of the overall architecture. For example, the robot was prepared to perform an *object recognition* task (Section 3.2) which tested the vision system and its short- and long-term memories in conjunction with natural language processing . The robot was asked questions about objects in its visual field (e.g., "What is this?", "How many coke cans do you see?"); if an object was unknown, the robot was told what it was and could successfully re-identify it at a later time. Another relatively basic task demonstrated on the robot was the *take orders* task, where the robot was able to perform a limited set of actions to confirm operability of the natural language processing, localization, and motor control (e.g., "Move forward 2 meters", "Turn left").

More ambitious was the *Open Interaction* task (also described in Section 3.2), which tested system cohesion and performance. In this task, the robot wandered an open area, detected nearby people, approached them, and initiated conversations.

The *waiter* task further explored the functionality of the entire system, with a focus on high-level task and environment knowledge, in addition to cognitive scripting capabilities. In particular, a set of scripts encoding typical interactions in a waiter/patron scenario were defined and carried out successfully by the robot (e.g., approaching a patron, taking a drink order, going to the "bar" to get the drink, and delivering the drink to the patron).

A final example, which has been evaluated formally in human subject experiments in our laboratory, is the *extra-planetary exploration* task (Scheutz et al., 2006), where a human and a robot must act jointly as a team in a fictitious planetary surface exploration scenario to accomplish the goal of finding an appropriate location on the planet from which to transmit geological data back to an orbiting spacecraft. This team task not only demonstrated that DIARC can be successfully employed in natural human-robot interactions, but also showed that affect expression – if employed correctly – can improve the performance of human-robot teams.

We will start with a brief summary of our findings about the utility of affect for NHL-HRI in the laboratory and then report the results from running DIARC on the robot in the unconstrained environment of the AAAI 2005 Robot Competition.

3.1 The Utility of Affect for NHL-HRI

Over the last decade, the potential of "affective computing" (Picard, 1997) prompted sub-communities in HCI, AI, and robotics to investigate useful roles of affect in artificial systems. There was already some early recognition of the potential utility of affective control for influencing the behavior of people (e.g., Breazeal & Scassellati, 1999). Moreover, studies with robots and simulated agents showed that emotional mechanisms can improve the performance of agents and may be cheaper than other, more complex non-emotional control mechanisms (e.g., Murphy, Lisetti, Tardif, Irish, & Gage, 2002; Scheutz & Logan, 2001). Even though many important advances have been made in our understanding of how to make machines recognize or signal different kinds of affect in interactions with people (e.g., see Rani, Sarkar, , & Smith, 2003; Lisetti, Brown, Alvarez, , & Marpaung, 2004; Kanda, Iwase, Shiomi, & Ishiguro, 2005), there is currently only one study (Scheutz et al., 2006) that investigated the effect of the robot's affect expression on team performance in a joint human-robot team task based on an *objective performance measure* ("time-to-task-completion").

In this study, human subjects were told that they had to find an appropriate transmission location (by directing the robot using natural language commands like "go forward", "turn right", "now take a reading") in the environment where the robot could transmit the "geological data" already stored in its memory. After one minute, stress was induced in subjects by virtue of a warning message uttered by the robot: "I just noticed that my battery level is somewhat low, <name>, we have to hurry up." A similar message was repeated after the second minute and, if the transmission site had not been located, the task ended after three minutes ("My batteries are dying, <name>. We have failed!"). Performance was measured in terms of the *time-to-task-competition*, and pre- and post-experiment surveys were conducted to ask subjects various questions about their perceptions of the robot.

Experiments were conducted with 50 subjects in two conditions: an *affect condition* where the robot's voice was modulated to express elevated stress starting with the first battery warning, and again to express even more stress at the second battery warning, and a *no-affect control condition* in which the robot's voice remained the same (see Scheutz et al., 2006 for details about the experimental setup). The results reported in Scheutz et al. (2006) show that subjects in the *affect condition* are overall faster in finishing the task than subjects in the *no-affect condition*, thus supporting the view that affect expression can have an objectively measurable, facilitatory effect. Here we extend the analysis in Scheutz et al. (2006) and examine both the subjects' own self-reported stress as well as the subjects' perceptions of robot stress based on their answers to questions on the post-experiment survey. First and foremost, we found a highly significant difference between pre- and post-announcement stress levels in all subjects ($t(98) = 5.59, p < .0001$), indicating a strong tendency for increased levels of stress in subjects after the first battery warning. While there was no difference among subjects in the two conditions with respect to pre-announcement stress ($t(48) = .74, p = .46$), the affect groups became more stressed than the no-affect groups after the battery warning, as indicated by a marginally significant difference in post-announcement stress ($t(48) = 1.82, p = .075$). Moreover, subjects in the *affect condition* are on average in agreement that the robot's stress levels had increased after it had issued the battery announcement, while subjects in the *no-affect condition* did not think that the robot was stressed, as indicated by a significant difference in perceived robot stress ($t(46) = 2.76, p = .008$). We also found a significant positive correlation of $r = .6$ between post-announcement stress and perceived robot stress ($t(1,19) = 3.33, p < .005$), which thus explains over a third of the variance in the no-affect groups ($R^2 = .37$). In contrast, no significant correlation was found in the affect groups. This suggests that the extent to which subjects in the no-affect groups perceived the robot as being stressed or not stressed depended in part on their own self-perceived stress (projected onto the robot), while in the affect groups the robot's perceived stress level likely depended (at least in part) on the robot's affective voice modulation.

The results then suggest that the content of the message was insufficient, in itself, to trigger in the no-affect subjects the belief that the robot might be "stressed," even though the robot was in exactly the same internal (stress) state with respect to its goal priorities and deadlines as in the affect conditions (the only architectural difference between the two conditions was that affective modulation of the robot's voice based on its internal states was suppressed in the no-affective condition). Rather, the affect modulation of the robot's voice seems to have contributed to the attribution of stress to the robot by subjects in the affect condition.

In sum, we believe that the results based on objective and subject evaluations demonstrate that appropriate affect expression (that is congruent with people's own affective states) can help humans in construing a mental model of a robot, which makes the robot's "mental states" transparent to the human and its behavior predictable. Such mental models might motivate subjects to help the robot (e.g., if they think it is stressed and maybe overwhelmed) or to try harder at achieving a task. Moreover, they might become more aware of the way they interact with the robot and automatically adapt their interaction patterns so as to facilitate interactions and performance, as suggested by our results.

3.2 The Robot at AAAI in 2005

The experiments above demonstrate the robot's competence under controlled laboratory conditions, in particular with respect to category (1). However, this competence does not necessarily translate directly to competence in the real world, where it is impossible to control all potentially confounding variables. Hence, it is important to verify the robot's performance in unstructured environments, such as the AAAI 2005 robot competition. Our entry (ND-Rudy) was prepared for two competition categories: the *Open Interaction Event* and the *Robot Exhibition*. The *Open Interaction* configuration would cause the robot to approach people and attempt to initiate simple template-based conversations (e.g., "Hello, my name is Rudy. Would you like to chat?"). A limit was placed on the length of a conversation, although the robot could terminate the conversation early if it noticed the human had not responded in a while.

Several demonstrations of the robot's abilities were prepared for the *Robot Exhibition*. To demonstrate visual short-term memory, the robot was asked to recall (without looking) how many faces were immediately surrounding it. The ability to identify objects visually was demonstrated, including the ability to learn the names of previously unknown object types and subsequently identify them correctly.

Each of these capabilities was implemented and tested in the lab prior to the competition. Although not perfect, the robot performed reasonably well in that controlled environment. It would wander and find people, engage in conversation, and then move on, and it would execute its demonstration tasks fairly consistently. However, the practical realities of the conference site proved more troublesome than anticipated.

Open Interaction. The large crowd (and attendant conversational hubbub) was problematic for the system, as were many physical features of the exhibition area (e.g., mirrors and tablecloths). The robot traversed the open interaction environment fairly well, but was unable to consistently engage people in conversations.

The lack of *automatic failure recovery* mechanisms in the infrastructure led at times to extended periods of inactivity while manual recovery was performed. Although ADE did include facilities to allow the user to restart individual components without bringing the entire system down, the lack of reliable wireless access in the competition area made it impossible to connect remotely, necessitating time-consuming system restarts.[4]

During the open interaction, the vision system was challenged by several factors. While the vision system worked well if run alone, running other concurrent compute-intensive processes negatively impacted performance. Lighting conditions were also problematic. One aim of the employed swarm mechanisms as part of the vision system was to adapt gracefully to changing light, however, the swarm system was in its early stages at the time of the competition. As such, feature extraction was less reliable than anticipated, leading to sometimes sporadic performance by components dependent on this output.

The robot was able to detect legs using the laser, and used them as a first pass in detecting humans during the open interaction; however, it was often the case that other "leg-like" structures (e.g., the ruffles of a tablecloth) would be identified as legs. Moreover, the leg detection did not distinguish the fronts of legs from the backs of legs, leading the robot to "initiate a conversation" with a human's back at times. Overall, we found that the people detection system placed too much weight on the laser evidence relative to face detection in concluding that a person was present and facing the robot.

The difficulties experienced in natural language processing were attributable mostly to the ambient noise level in the exhibition hall. Although Sonic performed reasonably well in the controlled environment of the lab, it was probably not optimally configured. Given the noise of the crowd at the competition, therefore, speech recognition was *very* limited. Attempts to improve performance by

[4] Automatic failure recovery procedures have since been added to the infrastructure to address this problem; see also Section 4.

reducing the size of the dictionary had only limited effect and were insufficient to allow Sonic to reliably recognize speech. Because all interactions were designed to be conducted via spoken natural language, this was clearly the most significant limitation of the robot. In the open interaction, the robot often seemed confused, responding inappropriately to what was said and sometimes prematurely ending a conversation because it mistakenly took the lack of intelligible output from Sonic as silence (and disinterest) on the part of the person.

Exhibition Demonstrations. Although some demonstrations completed successfully during the exhibition, there was insufficient time to perform others because of delays due to software failures. Moreover, the crowd noise was often so high that judges were unable to hear the robot's speech.

The ADE framework proved highly successful for distributing components across hosts; at the competition, various components were relocated to different computers in an effort to optimize overall performance. However, the uncontrolled environment exposed limitations in the visual object recognition system. The SIFT mechanism is susceptible to "pollution" by keypoints that are actually part of the background instead of the object being identified. It proved virtually impossible to avoid this pollution in real world scenarios, resulting in a number of misidentifications.[5]

Given the language understanding subsystem's inability to cope with the noise, it is unsurprising that the robot had trouble in many cases responding with correct behavior. Moreover, the ambient noise level also impacted the other side of language processing—speech production—often making it impossible to understand the speech output due to underpowered speakers. At one point during the exhibition judging we were forced to fall back to using the keyboard of an attached laptop for (typed) natural language input and its LCD for (printed) natural language output.

4 Discussion and Related Work

In the laboratory environment, at least, DIARC addresses the first and the third requirements set forth in Section 2, although the performance of many components could be improved. The robot is – to some extent – able to interact via natural language, detect emotions, and produce nonverbal cues, such as shifting gaze to indicate focus of attention. Moreover, the action interpreter allows the script designer to specify failure recovery mechanisms for a variety of contingencies, ranging from the very generic to the specific. The ADE framework provides mechanisms to allow multiple disparate components to be combined into a functional robotic system, and its failure recovery capabilities, although somewhat lacking at the time of the competition, have been greatly improved, thereby significantly increasing the system's robustness.

The problems experienced when moving from the controlled environment of the laboratory to the real-world environment of the robot competition were primarily in three areas: infrastructure, vision, and language processing. The main limitation of the infrastructure was the lack of automatic failure recovery. This has been addressed in the meantime via the inclusion of failure detection mechanisms that allow ADE to notice when a component has failed, a resource specification scheme by which it can determine whether there is a target machine matching the resource needs of the failed component, and a recovery mechanism that allows ADE to restart the failed component. The system has been expanded to include mechanisms for reasoning about the state of the system, allowing ADE to make intelligent decisions concerning the placement of system components (Kramer & Scheutz, 2006a).

The vision subsystem lacked the robustness required for effective operation in uncontrolled environments, making feature detection unreliable. The system's performance has since been enhanced substantially by extending the swarm systems to allow for hierarchical swarms that can track features at different levels of granularity and in different parameter spaces.

Language processing was by far the biggest problem encountered at the competition, because it comprises the main interface between the robot and the humans in its environment. Sonic has been replaced by Sphinx 4 as the main speech recognition platform used by the system, improving recognition rates somewhat. Moreover, new techniques for dynamically narrowing the dictionary to the target task when the domain is sufficiently limited have also improved performance. However, these mechanisms are not practical for tasks like the open interaction, where it can be difficult to predict the subject of conversation. We are exploring different hardware configurations to improve performance (e.g., a noise cancelling microphone array and a wireless microphone headset). However, speech recognition remains a major bottleneck for overall system performance.

Other competition participants are also pursuing goals that relate directly to the requirements above. The UML Robotics Lab, in investigating HRI for teleoperated robotics, have developed a robust infrastructure using *sliding scale autonomy* by which the robot can accede some control to the remote user when it does not know how to proceed; the robot continues to operate in

[5] This problem has since been addressed by applying a stereo vision algorithm to isolate an object in the foreground. Preprocessing the stereo image increases the likelihood that any of the determined keypoints actually belong to the object held to the camera.

some reduced capacity instead of stopping and waiting for human intervention (Desai & Yanco, 2005). Hansen Robotics is developing a realistic animated face that is capable of generating subtle visual cues (Hanson et al., 2005). The LABORIUS project confronts a substantial portion of the requirements in various subprojects. For natural language processing, they have developed a system for separating voices in order to understand multiple sentences simultaneously (Yamamoto et al., 2005). Their architecture for socially interactive robots includes motivational (although not explicitly affective) states that influence behavior (Michaud et al., 2005). And their Tito project implements nonverbal communication through the use of arm gestures, head shaking, and affect expression via smiling and raising eyebrows, and has the ability to recognize nonverbal cues, such as gaze detection, in humans (Michaud, Duquette, & Nadeau, 2003).

There are also other groups not represented at AAAI 2005 that are working on HRI projects similar to ours, using affect mechanisms to improve interactions. Here we can only review the two closest architectures in terms of using emotions for internal state changes and action selection. Murphy et al. (2002) implement emotional states with fixed associated action tendencies in a service robot as a function of two time parameters ("time-to-refill" and "time-to-empty") plus two constants. Effectively, emotion labels are associated with different intervals and cause state transitions in a Moore machine, which produces behaviors directly based on perceptions and emotional states. This is similar to the way *urgency* is calculated in our action manager, but different from the explicit goal representation used in our architecture, which allows for the explicit computation of the *importance* of a goal to the robot (based on positive and negative affective state), which in turn influences action selection (e.g., urgency alone may or may not result in reprioritization of goals and thus changes in affective state). Moreover, the robots in Murphy et al., 2002 do not use (spoken) natural language to interact with humans nor do they detect human affect.

The architecture in Breazeal, Hoffman, and Lockerd (2004) extends prior work (Breazeal, 2002) to include natural language processing and some higher level deliberative functions, most importantly, an implementation of "joint intention theory" that allows the robot to respond to human commands with gestures indicating a new focus of attention, etc. The system is intended to study collaboration and learning of joint tasks. One difference is that our robot lacks the ability to produce gestures beyond simple nodding and shaking by the pan-tilt unit (although it is mobile and fully autonomous as opposed to the robot in Breazeal et al., 2004). More importantly, the mechanisms for selecting subgoals, subscripts, and updating priorities of goals seem different in our affective action interpreter, which uses a dual representation of positive and negative affect that is influenced by various components in the architecture and used for the calculation of the importance, and consequently the priority, of goals.[6]

Despite significant advances in several areas of HRI, none of the systems that competed in 2005 (including ours) has demonstrated the second requirement for natural interaction with humans in real-world environments: the ability to demonstrate and recognize intent. It is in this general area that human-robot interaction is most lacking. As humans, we take for granted many of the subtle (even subliminal) cues present in any human-human interaction. The ability to integrate non-verbal information with explicitly communicated information (such as sarcasm or impatience) comes very naturally for humans; it can, in fact, be very difficult to specify what led one to a particular conclusion regarding another person's intent when it was not explicit in the spoken word. Similarly, humans instantly and often subconsciously make inferences of others' private mental states based on external clues. When one member of a group shares information, we infer immediately that all group members now have that new knowledge. Also, we are often able to make useful inferences based on what someone *does not* say (e.g., when the situation clearly requires some comment, but none is offered). These are only a few simple examples of an array of "rules of thumb" that humans use *and expect their interlocutors to use* to infer intent in the course of normal conversations. Without a systematic approach that exposes and incorporates this type of implicit knowledge, robots will continue to miss critical, if not constitutive, parts of human social interactions; hence, the second requirement will remain a road block to NHL-HRI, even if all the other problems are solved.

We believe that affect will play a pivotal role in removing this road block. The intentionality requirement is the furthest from fulfillment in part because of the tremendous complexity of goal-driven cognitive architectures. However, as we argue above, one function of affect is to integrate and manage many diverse cognitive processes, eliminating the need for complex centralized control. The result is an architecture in which multiple states (both internal and external) influence the prioritization and selection of goals, leading to predictable intentional behavior that humans can use to develop a "theory of mind" for the robot. DIARC is novel in this regard; affective states reflective of past experiences influence the operation of the goal-based control system. The system responds to events in a more "human-like" manner than non-affective systems, allowing humans to "relate" better and making them more likely to ascribe the property of intentionality to the robot. As such, DIARC is the among the most

[6] The details for reprioritization of goals were not provided in Breazeal et al. (2004).

advanced architectures available for NHL-HRI.

5 Conclusion

In this paper, we have proposed three main categories of requirements for natural human-like human-robot interaction: (1) social interaction abilities, such as natural-language production and understanding, situational knowledge, and expression and recognition of affect and other non-verbal cues; (2) goal-oriented cognition, which requires the robot to act in a purpose-driven manner, allowing humans to predict the robot's behaviors based on ascribed beliefs, intentions, and desires; and (3) robust intelligence, the ability to recover from failures within the system as well as failures of the system itself. The DIARC architecture introduced in the AAAI 2005 robot competition is a first attempt at meeting some of these requirements. While DIARC– and every other current robotic system – fails at achieving, even in part, the second requirement for NHL-HRI (intentionality) in natural environments, it does make substantial progress toward integrated social behaviors and fault tolerant cognition; it is able to recognize and express affect at a coarse-grained level, in addition to its natural language understanding and production capabilities, and ADE's automated failure recovery mechanisms greatly enhance the autonomy of the robotic system. Although the system proved too fragile to demonstrate these abilities in the real-world environment of the 2005 AAAI Robot Competition, its performance in experiments conducted with human subjects in a controlled laboratory setting is very promising for categories (1) and (3). Moreover, the degree to which affect is integrated into the architecture makes DIARC also a promising platform towards meeting the intentionality requirement (2) for NHL-HRI as the results from human subject experiments demonstrate.

Acknowledgment

The authors would like to thank Virgil Andronache, Chris Middendorff, Aaron Dingler, Peter Bui and Patrick Davis for their help with the implementation.

References

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

Andronache, V., & Scheutz, M. (2006). ADE - a tool for the development of distributed architectures for virtual and robotic agents. In P. Petta & J. Müller (Eds.), *Best of at2ai-4* (Vol. 20).

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.

Breazeal, C., Hoffman, G., & Lockerd, A. (2004). Teaching and working with robots as a collaboration. In *Proceedings of aamas 2004.*

Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. In *Iros* (pp. 858–863).

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

Burkhart, F. (2005). Emofilt: the simulation of emotional speech by prosody-transformation. In *Proceedings of interspeech 2005.*

Byers, Z., Dixon, M., Goodier, K., Grimm, C. M., & Smart, W. D. (2003). An autonomous robot photographer. In *Proceedings of iros 2003.* Las Vegas, NV.

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.

The CMU Sphinx group open source speech recognition engines. (2004). http://cmusphinx.sourceforge.net/html/cmusphinx.php.

Desai, M., & Yanco, H. A. (2005, August). Blending human and robot inputs for sliding scale autonomy. In *Proceedings of the 14th ieee international workshop on robot and human interactive communication.* Nashville, TN.

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

Ekman, P., & Friesen, W. V. (1977). *Manual for the facial action coding system (facs).* Palo Alto: Consulting Psychologists Press.

The festival speech synthesis system. (2004). http://www.cstr.ed.ac.uk/projects/festival/. Centre for Speech Technology Research.

Fillmore, C. J., Baker, C. F., & Sato, H. (2002). The framenet database and software tools. In *Proceedings of the third international conference on language resources and evaluation (lrec)* (pp. 1157–1160). Las Palmas, Spain.

Grosz, B. L., & Sidner, C. L. (1990). Plans for discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 417–444). MA: MIT Press.

Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., Fernandez, R., & Stephanou, H. (2005, July). Upending the uncanny valley. In *Pro-*

ceedings of the aaai-05 robot workshop. Pittsburgh, PA.

Ichise, R., Shapiro, D., & Langley, P. (2002). Learning hierarchical skills from observation. In *Proceedings of the fifth international conference on discovery science* (pp. 247–258).

Kanda, T., Iwase, K., Shiomi, M., & Ishiguro, H. (2005). A tension-moderating mechanism for promoting speech-based human-robot interaction. In *Iros* (p. 527-532).

Kipper, K., Dang, H., & Palmer, M. (2005, July). Class-based construction of a verb lexicon. In *Proceedings of AAAI 2000.*

Kramer, J., & Scheutz, M. (2006a). *Reflection and reasoning for system integrity in an agent architecture infrastructure.* (Under review)

Kramer, J., & Scheutz, M. (2006b). *ADE: A framework for robust complex robotic architectures.* (Under review)

Lisetti, C. L., Brown, S., Alvarez, K., , & Marpaung, A. (2004). A social informatics approach to human-robot interaction with an office service robot. *IEEE Transactions on Systems, Man, and Cybernetics–Special Issue on Human Robot Interaction*, *34*(2), 195–209.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91-110.

Michaud, F., Brosseau, Y., Côté, C., Létourneau, D., Moisan, P., Ponchon, A., Ra ievsky, C., Valin, J., Beaudry, É., & Kabanza, F. (2005). Modularity and integration in the design of a socially interactive robot. In *Proceedings IEEE international workshop on robot and human interactive communication* (pp. 172–177).

Michaud, F., Duquette, A., & Nadeau, I. (2003). Characteristics of mobile robotic toys for children with pervasive developmental disorders. In *Proceedings ieee conference on systems, man, and cybernetics.*

Middendorff, C., & Scheutz, M. (2006, June). Real-time evolving swarms for rapid pattern detection and tracking. In *Proceedings of artificial life x.*

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.

Ortony, A., Norman, D., & Revelle, W. (2005). Effective functioning: A three level model of affect, motivation, cognition, and behavior. In J. Fellous & M. Arbib (Eds.), *Who needs emotions? the brain meets the machine.* New York: Oxford University Press.

Pellom, B., & Hacioglu, K. (2003). Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP).*

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

Rani, P., Sarkar, N., , & Smith, C. A. (2003). Affect-sensitive human-robot cooperation-theory and experiments. In *Proceedings of ieee international conference on robotics and automation (icra)* (pp. 2382–2387).

Schank, R., & Abelson, R. R. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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, quebec, canada, may 14-17, 2000, proceedings* (Vol. 1822, pp. 389–399). Springer.

Scheutz, M. (2004). Useful roles of emotions in artificial agents: A case study from artificial life. In *Proceedings of aaai 2004.*

Scheutz, M. (2006). ADE - steps towards a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence*, *20*(4-5).

Scheutz, M., & Andronache, V. (2003). APOC - a framework for complex agents. In *Proceedings of the aaai spring symposium.* AAAI Press.

Scheutz, M., & Andronache, V. (2004). Architectural mechanisms for dynamic changes of behavior selection strategies in behavior-based systems. *IEEE Transactions of System, Man, and Cybernetics Part B*, *34*(6), 2377-2395.

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., McRaven, J., & Cserey, G. (2004). Fast, reliable, adaptive, bimodal people tracking for indoor environments. In *IEEE/RSJ international conference on intelligent robots and systems (IROS).*

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 ACM conference on human-robot interaction (HRI2006).*

Sleator, D., & Temperley, D. (1993). Parsing english with a link grammar. In *Proceedings of the third international workshop on parsing technologies.*

Yamamoto, S., Nakadai, K., Valin, J., Rouat, J., Michaud, F., Komatani, K., Ogata, T., & Okuno, H. (2005). Making a robot recognize three simultaneous sentences in real-time. In *Proceedings IEEE/RSJ international conference on intelligent robots and systems* (pp. 897–902).

Yang, M.-H., Kriegman, D. J., & Ahuja, N. (2002, January). Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(1), 34–58.