Modeling and evaluating empathy in embodied companion agents

Scott W. McQuiggan, James C. Lester

Dept. of Computer Science, 890 Oval Drive,
North Carolina State University
Raleigh, NC 27695, USA

{swmcquig, lester}@ncsu.edu

Abstract

Affective reasoning plays an increasingly important role in cognitive accounts of social interaction. Humans continuously assess one another's situational context, modify their own affective state accordingly, and then respond to these outcomes by expressing empathetic behaviors. Synthetic agents serving as companions should respond similarly. However, empathetic reasoning is riddled with the complexities stemming from the myriad factors bearing upon situational assessment. A key challenge posed by affective reasoning in synthetic agents is devising empirically informed models of empathy that accurately respond in social situations. This paper presents CARE, a data-driven affective architecture and methodology for learning models of empathy by observing human-human social interactions. First, in CARE training sessions, one trainer directs synthetic agents to perform a sequence of tasks while another trainer manipulates companion agents' affective states to produce empathetic behaviors (spoken language, gesture, and posture). CARE tracks situational data including locational, intentional, and temporal information to induce a model of empathy. At runtime, CARE uses the model of empathy to drive situation-appropriate empathetic behaviors. CARE has been used in a virtual environment testbed. Two complementary studies investigating the predictive accuracy and perceived accuracy of CARE-induced models of empathy suggest that the CARE paradigm can provide the basis for effective empathetic behavior control in embodied companion agents.

1. Introduction

There is a growing demand for interactive technologies to create engaging experiences for increasingly sophisticated users. In response, recent years have witnessed significant progress on synthetic agents inhabiting interactive systems with a broad range of applications in entertainment, education, and training. Foundational work on synthetic

agents has yielded expressive models of embodied cognition and behavior that support rich interactions in virtual environments (André and Müller, 2003; Bates, 1994; Cavazza et al., 2002; Johnson and Rickel, 1998; Lester et al., 2000; Swartout et al., 2004). Complementing advances in cognition and behavior, affective reasoning (Elliott, 1992; Gratch and Marsella, 2004; Ortony et al., 1988; Picard, 1997; Porayaska-Pomsta and Pain, 2004) has begun to play a central role in human-computer interaction (Hudlicka, 2003) and the design of synthetic agents (Bates et al., 1992; Bickmore, 2003; Burleson and Picard, 2004; Marsella and Gratch, 2003) and embodied conversational agents (André et al., 2000; Cassell et al., 2000; de Rosis et al., 2003; Lester et al., 2000; Nass et al., 2000; Rickel and Johnson, 2000). The community is now well positioned to investigate affective reasoning in the context of social interaction (Brave et al., 2005; Johnson and Rizzo, 2004; Paiva et al., 2005; Prendinger and Ishizuka, 2005). Transitioning affective synthetic agents into the social arena could yield companion agents that provide users with motivating support and comfort. *Companion agents* can facilitate social interaction, a critical capability in virtual environments for education (Burleson and Picard, 2004; Conati, 2002; Lester et al., 2000) and training (Prendinger and Ishizuka, 2005). Companion agents help users cope with frustration (Burleson and Picard, 2004), deal with stress (Prendinger and Ishizuka, 2005), and counsel children on social behaviors, such as bullying in schools (Paiva et al., 2005).

Empathy is a key component of social interaction (Hoffman, 2000). Because empathetic companion agents hold much promise for socially engaging virtual environments, empathy modeling is a logical next step in the evolution of synthetic agents. One can distinguish two fundamental approaches to modeling empathy: analytical and empirical. In the *analytical* approach, models of empathy can be constructed by analyzing the findings of the empathy literature. However, empathy is not well understood. It is only in the past two decades—this is very recent in the history of psychology—that empathy has become a focus of study for social psychologists (Davis, 1994). Perhaps as a result of its limited study, while we have expressive computational models of affect, e.g., the OCC model (Ortony et al., 1988), we do not have similarly rich models of empathy. Moreover, because empathetic reasoning requires drawing inferences about another's intentions, her affective state, and her situational context, devising a universal model of empathy seems to be well beyond our grasp at the current juncture.

An alternative to analytically devising models of empathy for synthetic agents is the *empirical* approach. If somehow we could create models of empathy that were derived directly from observations of "empathy in action," we could create empirically grounded models based on human-human empathetic behaviors exhibited during the performance of a specific task within a given domain. While it is not apparent that this approach could produce a universal model of empathy—a universal model may not even be achievable, at least in the near term—the empirical approach could nonetheless generate models of empathy that significantly extend the communicative capabilities of socially intelligent agents.

The empirical approach calls for a data-driven framework for modeling empathy. This paper presents CARE,[1] a data-driven affective architecture and methodology for learning

[1] CARE: Companion-Assisted Reactive Empathizer.

empirically informed models of empathy from observations of human-human social interactions. During training sessions, CARE monitors situational data including locational, intentional, and temporal information while one trainer (the *target*) directs her agent to perform a sequence of tasks in a virtual environment as another trainer (the *empathizer*) reactively manipulates her agent's affective state to produce empathetic behaviors (spoken language, gesture, and posture). Inducing a model of empathy, CARE uses situational data as predictive features for empathy assessment (when to exhibit an empathetic behavior) and for empathy interpretation (which levels of valence and arousal should be chosen, i.e., the affective state). At runtime, CARE uses the resulting model to drive situation-appropriate empathetic behaviors in the companion agent as it interacts with actual users.

Empathetic accuracy is the accuracy with which an empathizer in a social context assesses another's thoughts and feelings and then acts empathetically (Ickes, 1997). The empathetic accuracy of a model of empathy can be determined with two complementary types of evaluations:

- *Predictive Accuracy*: Using a *k*-fold cross validation approach commonly used in the machine learning community, a predictive accuracy study investigates the empathetic accuracy of a model by determining the predictive accuracy of the model relative to the empathetic decisions made by humans in similar social contexts. A predictive accuracy study can reveal the degree to which a model of empathy makes assessment and interpretation decisions that accurately emulate humans' assessment and interpretation decisions.
- *Perceived Accuracy*: A perceived accuracy study investigates the empathetic accuracy of a model with a controlled focus group experiment. Competing empathy models are incorporated into multiple embodied companion agents, subjects observe the companion agents in a range of social contexts, and the subjects rate the situational appropriateness of the agents' empathetic behaviors. A perceived accuracy study can reveal the degree to which a model of empathy makes assessment and interpretation decisions that are perceived by humans to be situationally appropriate.

CARE models have been evaluated in a predictive accuracy study and a perceived accuracy study. Each study involved 31 subjects – there was no overlap of subjects in the two studies – and results indicate that CARE models generate empathetic behaviors that are similar to those made by humans and are perceived to be situationally appropriate.

The article is organized as follows: Section 2 provides background on empathy and affective reasoning in synthetic agents. Section 3 presents the CARE architecture and methodology. CARE has been used to create a model of empathy for an embodied companion agent inhabiting Treasure Hunt (Fig. 1), a virtual environment in which a user and a companion agent search for treasures. Section 4 describes the CARE implementation and its generation of models of empathy in the Treasure Hunt companion agent. Section 5 presents a predictive accuracy evaluation of CARE-induced models of empathy. Section 6 presents a perceived accuracy evaluation of CARE empathy models. Concluding remarks and directions for future work follow in Section 7.

Fig. 1. Treasure Hunt world with companion agent (left) and the target's agent (right).

2. Empathy

Devising computational models of empathy contributes to the broader enterprise of modeling affective reasoning (Picard, 1997). Beginning with Elliott's implementation (Elliott, 1992) of the OCC model (Ortony et al., 1988), advances in affective reasoning have accelerated in the past few years, including the appearance of a sophisticated theory of appraisal (Gratch and Marsella, 2004) based on the Smith and Lazarus Appraisal Theory (Lazarus, 1991). We have also begun to see probabilistic approaches to assessing users' affective state in educational games (Conati, 2002) and investigations of the role of affect and social factors in pedagogical agents (Baylor, 2005; Burleson and Picard, 2004; Elliott et al., 1999; Johnson and Rizzo, 2004; Lester et al., 2000; Prendinger and Ishizuka, 2005). Recent work on empathy in synthetic agents has explored their affective responsiveness to biofeedback information and the communicative context (Prendinger and Ishizuka, 2005). It has also yielded agents that interact with one another and with the user in a virtual learning environment to elicit empathetic behaviors from its users (Paiva et al., 2005). Empathy has also been investigated in embodied computer agents perceived to care about outcomes of human user experiences in a blackjack game (Brave et al., 2005).

Empathy is a complex socio-psychological construct. Defined as "the cognitive awareness of another person's internal states, that is, his thoughts, feelings, perceptions, and intentions" (Ickes, 1997), empathy enables us to vicariously respond to another via "psychological processes that make a person have feelings that are more congruent with another's situation than with his own situation" (Hoffman, 2000).

Social psychologists describe three constituents of empathy. First, the *antecedent* consists of the empathizer's consideration of herself, the target's intent and affective state, and the situation at hand. Second, a*ssessment* consists of evaluating the antecedent. Third, *empathetic outcomes*, e.g., behaviors expressing concern, are the products of assessment (Davis, 1994) including both affective and non-affective outcomes (e.g.,

judgment, cognitive awareness). Two types of affective outcomes are possible. In *parallel outcomes*, the empathizer mimics the affective state of the target. For example, the empathizer may become fearful when assessing a target's situation in which the target is afraid. In *reactive outcomes*, empathizers exhibit a higher cognitive awareness of the situation to react with empathetic behaviors that do not necessarily match those of the target's affective state. For example, empathizers may become frustrated when the target does not meet with success in her task, even if the target herself may not be frustrated. Accurately modeling parallel and reactive empathetic reasoning presents significant challenges.

3. Data-Driven Evaluative Empathy Modeling

The prospect of creating an "empathy learner" that can induce empirically grounded models of empathy from observations of human-human social interactions holds much appeal. To this end, this article proposes CARE, an affective data-driven paradigm that learns empathetic assessment (when to be empathetic) and empathetic interpretation (how to be empathetic).

3.1. Architecture

The CARE architecture operates in two modes: empathetic model induction in which the architecture interacts with two trainers (depicted in the diagram with dotted arcs), and runtime operation, in which it manages empathetic behaviors for a companion agent interacting with a user (depicted in the diagram with solid arcs) (Fig. 2):

- **Empathetic Model Induction**: Trainers interact with CARE via interfaces through which they direct synthetic agents in the virtual environment. The virtual environment tracks all activities in the world and reports observable attributes pertaining to temporal, locational, and intentional information. These are passed to the empathy learner during the training phase. During the subsequent learning phase, the learner induces a model of empathy that is operational, i.e., it can be used at runtime.
- **Runtime Operation**: Users interact with CARE via an interface through which they direct a synthetic agent in the virtual environment. Throughout their experience, they interact with a companion agent controlled by CARE. The virtual environment again tracks all activities in the world and monitors the same observable attributes reported to the empathy learner during empathetic model induction. The induced model is used by the empathetic behavior manager to (1) assess the situation to determine *when* to be empathetic, and (2) interpret situations deemed "empathy-worthy" to decide *how* to be empathetic. When a situation calls for empathy, a suitable empathetic behavior (including speech, gesture, and posture) is selected for execution by the companion agent to react empathetically to the user's situation. Spoken components of companion agent empathetic behaviors explicitly state the affective state being conveyed, e.g., "This has become quite frustrating." Verbal communications are accompanied by commonly associated gestures and posture, i.e., *dropped shoulders*, *arms crossed*, *looking down*, and *head shaking*.

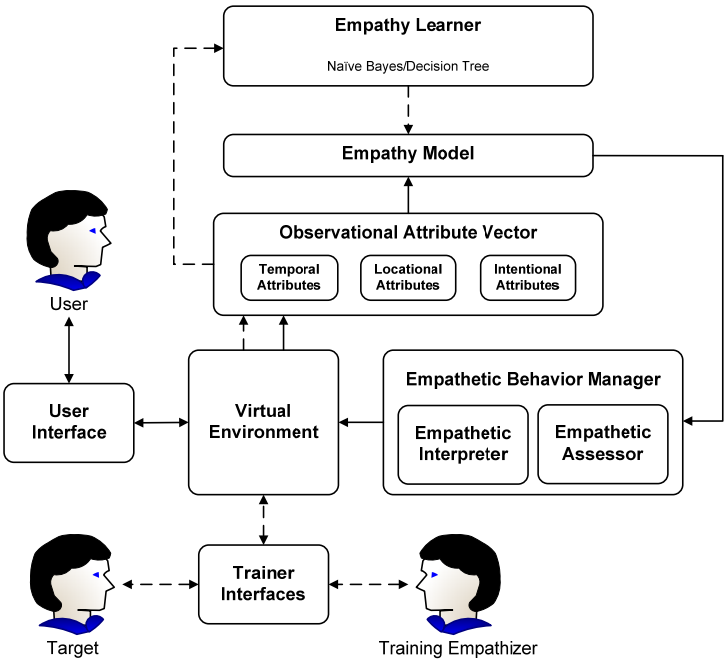

Fig. 2. CARE empathy modeling architecture.

3.2. Training and Learning

In the training phase, CARE's trainable agent must be exposed to social situations similar to the ones it will encounter at runtime. Because empathy by its very nature involves multiple actors (here we focus on two), the training experience should revolve around the interaction of multiple subjects in situations that elicit empathetic behaviors.

CARE training sessions are therefore situated in task-oriented scenarios involving two trainers, a *target* and an *empathizer*, each of whom is represented by a synthetic agent in the 3D virtual environment where training takes place. The target, whom is given a multi-objective mission to complete, controls her agent to navigate and perform tasks in the virtual environment from a first person point-of-view (POV). It is the task of the empathizer, who looks on from a third-person POV, to monitor the target's activities and select suitable empathetic affective states based on the target's observed behaviors. Selecting an affective state causes her agent to perform an empathetic behavior.

To collect empathy data that is as representative as possible of that which will be encountered by the companion agent at runtime, training sessions must satisfy the following requirements:

- ***Affective space coverage***: To promote the target's experiencing a range of emotions spanning the classic two-dimensional affective space defined by *valence* (degree of attraction, ranging from negative to positive) and *arousal* (level of stimulation, ranging from low to high) (Lang, 1995) (Fig. 3), the target should be faced with goals of varying degrees of difficulty: some should be very easy to achieve, while others should

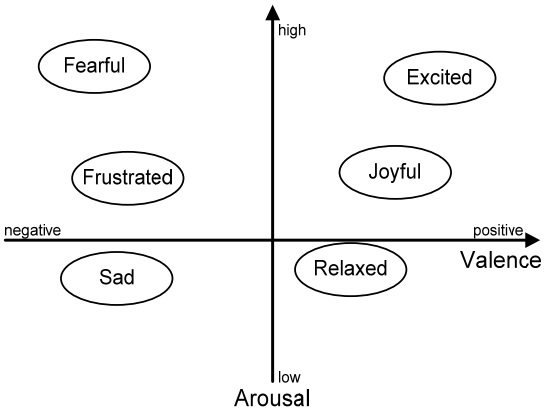

Fig. 3. Two-dimensional affective space.

be very challenging. For example, in Treasure Hunt, the virtual environment that serves as a testbed for CARE, some treasures are in plain view of the target while others are partially occluded; some are hidden altogether. Some targets should be exposed to virtual environments in which goals are easy to achieve, and some should be introduced into worlds in which goals are difficult to achieve. Thus, in some Treasure Hunt worlds, targets can score a specified number of points by collecting treasures very easily, while other worlds pose significant challenges stemming from the accessibility and varying point values. These unique situations offer opportunities for users to experience a variety of reactive emotions.

- ***Double-blind training***: Training sessions should be conducted in such a manner that the target is unaware that an empathizer is at the controls of the empathetic behaviors of the companion agent in the virtual world. Likewise, restricting the empathizer's environment to the virtual world (i.e., without access to the target's facial or vocal expressions) enables empathetic decisions to be based solely on inferences from the observed virtual world (thus, similar inferences are likely to be made by the empathy models at runtime).
- ***Controlled affective expression***: Minimizing the complexity of the empathizer's task can be achieved by limiting the set of emotions at her disposal. For example, empathizers in Treasure Hunt have access to six affective states: *excited*, *joyful*, *relaxed*, *fearful*, *frustrated*, and *sad*. This particular set of emotions was chosen because it covers the four quadrants of the two-dimensional affective space (Lang, 1995) and considers three levels arousal (high, medium and low) for each level of valence (positive or negative).
- ***Uniform agent personae***: While investigating different personae is a promising direction for future work, e.g., pedagogical agent personae experiments (Baylor, 2005), baseline training should control for personae by holding both the target's agent and the empathizer's agent constant throughout training sessions.
- ***Situation data collection intervals***: Situation data should be collected at least as often as significant events occur, where an event is deemed significant if it can plausibly affect the empathizer's decisions.

Accurately modeling empathy requires a representation of the situational context that satisfies two requirements. First, it must be sufficiently rich to support empathetic assessment and empathetic interpretation. Second, it must be encoded with features that are readily observable at runtime so that they may drive companion agents' empathetic decision making. CARE therefore employs an expressive representation of all activities in the virtual environment by encoding them in an observational attribute vector that is used in both modes of operation: during empathetic model induction, the *observational attribute vector* is passed to the empathy learner for model generation; during runtime operation, the attribute vector is monitored by the empathetic behavior manager for determining empathetic behavior. CARE's observable attribute vector represents three interrelated categories of features for making empathetic decisions:

- **Temporal features**: CARE tracks the amount of time that has elapsed since the target/user arrived at the current location, since the target/user achieved a goal, since the empathizer/companion agent last behaved empathetically, and since the target/user was last presented with an opportunity to achieve a goal.
- **Locational features**: CARE continuously tracks the location of all agents in the environment. It monitors locations visited in the past, locations recently visited, locations not visited, and locations being approached. Locations are associated with specific areas in the virtual environment or areas containing significant objects or obstacles, e.g., goals or locked doors.
- **Intentional features:** CARE tracks goals being attempted (as inferred from locational and temporal features, e.g., approaching a location where a goal can be achieved), quantity and quality of goals achieved, the rate of goal achievement, and the effort expended to achieve a goal (as inferred from recent exploratory activities and locational features).

In the CARE implementation for Treasure Hunt, the observational attribute vector encodes 192 features. During empathetic model induction, an instance of the vector is logged every time a significant event occurs. On average, vectors are updated several hundred times each minute. At runtime, the same features are updated continuously by the virtual environment and are used by the empathetic behavior manager to select situation-appropriate empathetic behaviors. Figure 4 shows how information from observations in CARE training sessions flows from the training phase to the learning phase for empathetic model induction.

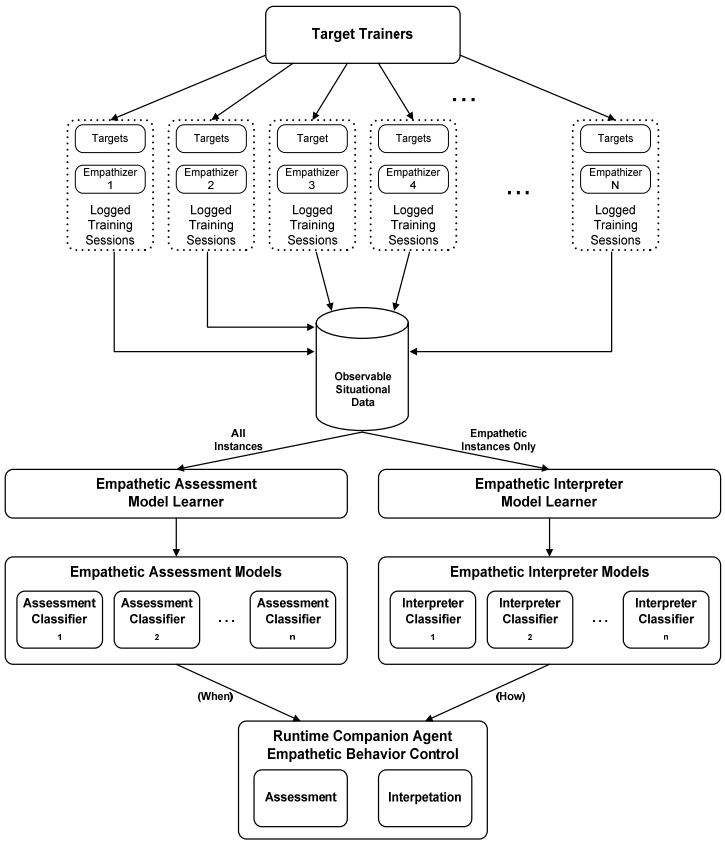

Fig. 4. CARE framework data flow.

Finally, in the learning phase, CARE induces a dual model of empathy. One component will be used at runtime to support empathetic assessment, and the other will be used to support empathetic interpretation. CARE's empathy learner first uses all of the data collected in the training session to induce the empathetic assessment model. Induction may be based on any standard classifier learning technique. Two versions of CARE have been implemented in Treasure Hunt, one with a naïve Bayes classifier and one with a decision tree classifier. The evaluation reported in Section 5 discusses the performance of both approaches. CARE's empathy learner next uses a subset of the data collected in the training sessions to induce the empathetic interpretation model. Here, it only considers data instances in which empathy was in fact exhibited. The second induction produces a model of empathy interpretation that at runtime is used to guide agent's empathetic behaviors. The products of the learning phase are two classifiers used to determine when and how the companion agent should be empathetic as dictated by a generalized model induced from all of the empathizing trainers' empathetic behavior decisions. Because the classifiers employ features directly observable in the environment, they can be easily integrated into the runtime behavior control systems of companion agents in the form of rules or probabilistic statements.

Fig. 5. A frustrated companion agent and target agent.

4. Treasure Hunt Prototype Virtual Environment

The CARE paradigm has been used to train models of empathy and to control the behavior of a companion agent at runtime in Treasure Hunt, a virtual environment testbed in which targets/users are instructed to collect as many treasures as they can in the allotted time.

4.1. Treasure Hunt

Treasure Hunt is a prototype virtual environment featuring a synthetic agent controlled by the user and a companion agent whose empathetic behaviors are controlled by CARE. The user navigates the 3D virtual world in search of hidden (and some not- so-hidden) treasures. Each treasure box is labeled with the value of its contents, representing points the user obtains when collecting the associated treasure. Some treasure boxes are cryptically labeled, hiding the value of its contents from users. Throughout the users' quest for treasure, the companion agent follows along and expresses empathetic behaviors as appropriate situations arise in the users' experiences (Fig. 5).

4.2. *Implementation*

CARE's empathetic assessment model and interpretation model have been implemented using naïve Bayes and decision tree approaches. A discussion of their relative performance follows in Section 5. The empathetic models were induced from a dataset consisting of a 192-dimensional observational attribute vector. The observational attribute vector consists of temporal, locational, and intentional features. For example, a sliding ten-second window was used as a temporal feature for tracking user goal attainment, while a binary locational feature monitored whether the user had yet visited

the docks or rocky beach area, and an intentional feature was used to detect when the user was moving in the direction of a high-valued goal in the her view.

Treasure Hunt was implemented using a high-performance 3D game platform from Valve Software.

4.3. Example Scenario

To illustrate the empathetic behavior control posed by CARE, consider the following scenario, which repeatedly played out in CARE training sessions. As we catch up with the user, she has navigated her synthetic agent throughout the virtual environment struggling to find significant, high-valued treasure. The user and empathizer are aware that the user has not yet met her expected treasure collection quota (as depicted in the graphical HUD representation in the bottom corner of the display) and is quickly running out of time. Only 30 seconds remain.

The user has found her way into a location on the beach of the Treasure Hunt virtual environment, a location visited by the user's agent early in the session. The empathizer realizes that this particular location has been previously visited and was already determined to be an area without any treasure boxes. It has now been over one minute since the user last discovered any treasure at all. Assessing the situation, the empathizer selects the frustrated affective state, thereby initiating a behavioral sequence in which the companion agent announces her frustration by directly stating, "This is becoming quite frustrating," and using gestures and posture similar to the companion agent depicted in Figure 5. (The agent's speech segments are stored in pre-rendered audio clips.)

CARE's empathy learner monitored a variety of environmental characteristics, including those described above, during its training sessions. These recorded instances aid the empathy models in reproducing similar appropriate inferences in analogous situations where time is running out, the user's agent is in a previously visited location known to be without treasure, and the user's intended treasure collection goal is likely to fail. Thus, given the same situation with CARE driving the empathetic behaviors of the companion agent at runtime, empathetic assessment and interpreter models are likely to make similar appropriate empathetic decisions.

5. Evaluating Empathy Models: Predictive Accuracy

Two complementary approaches can be taken to evaluating the empathetic accuracy (Ickes, 1997) of a model of empathy. First, the *predictive accuracy* of a model can be evaluated. The predictive accuracy of a model of empathy is the degree to which it makes assessment and interpretation decisions that accurately emulate those made by humans. Second, the *perceived accuracy* of a model can be evaluated. The perceived accuracy of a model of empathy is the degree to which it makes assessment and interpretation decisions that are perceived by humans to be situationally appropriate.

This section (Section 5) describes an evaluation that investigates the predictive accuracy of CARE-induced models of empathy.

5.1. Method

5.1.1. Participants and Design

In a formal evaluation, more than two hours of data were gathered from thirty-one subjects in an Institutional Review Board (IRB) of North Carolina State University approved user study. The subjects were divided into 25 targets and 6 training empathizers. There were 20 male subjects and 5 female subjects serving as targets. There were 3 male and 3 female subjects participating as training empathizers. Subjects varied in race, ethnicity, age and marital status. On average, each training empathizer completed 4 training sessions, each with a different target participant.

5.1.2. Materials and Apparatus – Training Target

For each target participant pre-experiment paper-and-pencil materials consisted of a demographic survey, Half-Life 2 controls reference sheet, and a controlled backstory in preparation for interacting within the environment. The post-experiment paper-and-pencil materials consisted of a general survey about the training target's experience and opinions on affect in applications such as games. The demographic survey collected basic information such as gender, age, ethnicity, marital status, and number of children. The Half-Life controls reference sheet described which keys and mouse movements would be needed to manipulate the agent in both the practice task and the training task. The controlled backstory for the interactive environment was constructed in such a way that each participant would be given the same preparatory information.

The computerized materials for the targets consisted of three 3D Treasure Hunt virtual environments, each of varying degrees of difficultly, and the practice task drawn directly from the game Half-Life 2. The easiest version of Treasure Hunt offered many opportunities to find treasures and meet the expectations that were set in the backstory. The most challenging version of Treasure Hunt made it difficult to find treasures; there were fewer treasures worth less value and more occluded treasure boxes making it difficult to meet backstory expectations. The practice task from the game Half-Life 2 presented an opportunity for targets to become familiar with the required controls. The practice task required completing activities such things as climbing a ladder, stacking boxes, and jumping.

5.1.3. Materials and Appartus - Empathizer

For each training empathizer, pre-experiment paper-and-pencil materials consisted of a demographic survey, Davis' Interpersonal Reactivity Index questionnaire (Davis, 1983), a two-page summary of emotion and empathy constructs, and an empathizer controls reference sheet. Post-experiment paper-and-pencil materials consisted of a survey inquiring about the emotions used/unused, other emotions that could have been useful, and general opinions regarding affect in applications, such as games.

Before empathizers began training, they completed Davis's Interpersonal Reactivity Index (IRI) to obtain a measure of their empathy. The IRI consists of 28 statements in which respondents are instructed to rate the degree to which each item describes them on

a Likert scale of 0 to 4. The result is a set of 4 subscale values pertaining to the following qualities of empathy: *fantasy*, *perspective taking*, *empathetic concern*, and *personal distress* (Davis, 1994).

The computerized materials consisted of a spectator view (third person point of view) of the 3D virtual environment, Treasure Hunt, in which target trainers would be interacting. Empathizers did not view target trainer practice tasks and they were not informed about the degree of difficulty.

5.2. Procedure

Each training target participant entered a conference room and was seated in front of a laptop computer. First, target participants completed the demographic survey at their own rate. Concurrently, empathizers entered a second room and were seated in front of another laptop computer. Targets were unaware of the empathizer's participation at this point. Empathizers were only aware that a target training participant was in the next room. To ensure that empathizers only had access to characteristics of the target participant that could be obtained from the virtual environment, there was no physical, visual, or audio contact between the target and empathizer participants at any point. Like targets, empathizers also first completed the same demographic survey. Next, empathizers completed Davis' IRI questionnaire, while targets were given the Half-Life 2 controls reference sheet to read until the practice task was loaded on the laptop in front of the target. Target trainers then completed the practice task at their own rate. At this point, empathizers were given the emotion and empathy reference sheet and instructed to familiarize themselves with the definitions and empathizer controls. Next, one of the degrees of difficulty was randomly selected and that Treasure Hunt training environment was loaded on the target machine while the spectator view application was concurrently loaded on the empathizer machine.

After the training environment was loaded, target trainers had 7 minutes to explore the environment and collect treasure. Empathizers viewed the interaction and made empathetic behavior decisions by selecting the appropriate control for the affective state they desired the companion agent to express. When empathetic behaviors were selected by the empathizer, both participants had the opportunity to hear the companion agent's spoken language and see the associated gestural behaviors and posture. Upon completion of the 7 minute training session, both training targets and empathizers were given post-session surveys and were interviewed. Finally, target trainers were offered information about the details of the experiment and informed about the presence of the empathizer during the training session.

The following procedure was used to generate models of empathy from the training sessions (Figure 5-1 presents the evaluation data flow):

1. *Data construction*. Each session log, containing 6,000 – 9,000 observation changes, was first translated into a full observational attribute vector. For example, if a treasure box came into view (and all other observable attributes remained constant) then the observational attribute vector would modify the previous vector to account for the noted change.
2. *Data cleansing*. After data was converted into the observational attribute vector format, the data was ready to be cleaned. This step included partitioning the dataset

into one set containing only records in which the empathizer selected empathetic actions and one in which she did not. [2]

3. *Naïve Bayes classifier and decision tree induction*. The resulting data were loaded into the Weka machine learning package (Witten and Frank, 2005), and a naïve Bayes classifier and a decision tree were learned.
4. *Cross-validation analyses*. Tenfold cross-validation analyses were run on the resulting naïve Bayes and decision tree models. While the entire dataset was used to generate models for empathetic assessment (when to be empathetic) and empathetic interpretation (how to be empathetic), empathetic interpretation is induced solely from data in which empathy is exhibited.

5.3. Results

Both naïve Bayes and decision tree models were induced from data collected in the training sessions described above. As noted earlier, 192 observational attributes were used to define the feature vectors. Naïve Bayes and decision tree classifiers are effective machine learning techniques for generating preliminary predictive models. Naïve Bayes classification approaches produce probability tables that can be incorporated into runtime systems and used to continually update probabilities for monitoring when and how to be empathetic. Decision trees provide interpretable rules that support runtime decision making. Both the naïve Bayes and decision tree machine learning classification techniques are useful for preliminary predictive model induction for large multidimensional data.

Both models were evaluated using the *k*-fold cross-validation methodology. In *k*-fold cross-validations, which are used to obtain arbitrarily accurate estimates of error (Witten and Frank, 2005), data is decomposed into equal partitions: all but one partition are used for training, and one is used for testing., In each "run," testing is performed only on testing data, not on data used to train the model. Over the course of the "runs," the equal parts are swapped between training and testing sets until each partition has been used for both training and testing. Following the standard approach of using a value of 10 for *k*, the analyses described here employ a 10-fold cross validation (Witten and Frank, 2005).

Cross-validated ROC curves are useful for presenting the performance of classification algorithms for two reasons. First, they represent the positive classifications (true positives), included in a sample, as a percentage of the total number of positive classifications along the vertical axis, against the negative classifications (false positives) as a percentage of the total number of negative classifications along the horizontal axis (Witten and Frank, 2005). Second, the area under ROC curves is widely accepted as a generalization of the measure of the probability of correctly classifying an instance (Hanley and McNeil, 1982).

Figure 6 shows the ROC curves for CARE's naïve Bayes and decision tree classification approaches for modeling empathetic assessment. The ROC curves for each model predicted empathetic behavior triggers in a ten second interval. The area under the naïve Bayes curve in Figure 6 is 0.72 and the area under the Decision Tree curve is 0.89.

---

[2] In later steps of the procedure, learning empathy assessment will use both data sets, while learning empathy interpretation will use only the data set containing empathetic actions.

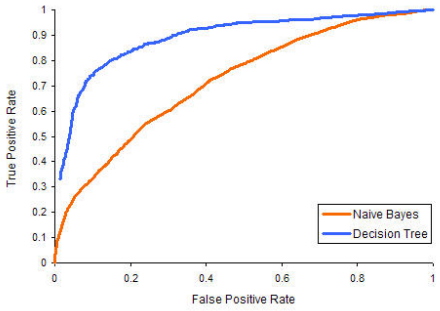

Fig. 6. Empathetic Assessment

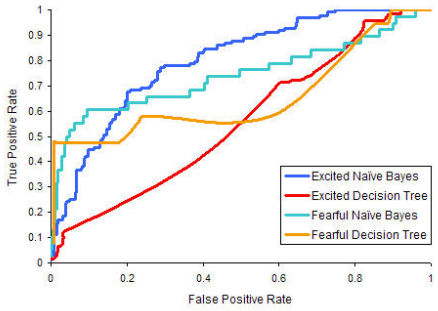

Fig. 7. Empathetic Interpretation

Figure 6 shows ROC curves for CARE's naïve Bayes and decision tree classification approaches for empathetic interpreter modeling. The ROC curves for naïve Bayes and decision tree models for empathetic interpretation for the affective states *excited* and *fearful* are found in Figure 7. Areas under each curve are as follows: 0.80 (Excited Naïve Bayes), 0.56 (Excited Decision Tree), 0.74 (Fearful Naïve Bayes), and 0.66 (Fearful Decision Tree).

5.4 Discussion

Two categories of functionality can be distinguished. First, the decision tree classifier was best suited for modeling empathy assessment, i.e., it was better able to determine *when* to be empathetic (Figure 6). Second, the naïve Bayes classifier was best suited to modeling empathy interpretation, i.e., it was better able to determine *how* to be empathetic (Figure 7). Although the figure shows only the excited and fearful emotions, all six emotions were evaluated and the naïve Bayes classifier bested the decision tree classifier in every case.

The smoothness of the curve in Figure 6 indicates that sufficient data seem to have been used for training empathy assessment, while the jaggedness of the curve in Figure 7 indicates that more data covering a large space of situations is called for in training empathy interpretation. For example, many empathizers only rarely used particular emotions, e.g., sad, and some trainers suggested that having more affective states available would have been helpful. In general, however, it appears that effective classifiers can indeed be learned for both empathy assessment and empathy interpretation.

Only 388 instances were available for modeling empathetic interpretation. Collecting more data would likely improve the predictability of the decision tree classifier for interpreting *how* to be empathetic. We speculate that for this reason, the decision tree classifier was outperformed by the naïve Bayes classifier for modeling empathetic interpretation.

6. Evaluating Empathy Models: Perceived Accuracy

Recall that *perceived accuracy* is the degree to which a model of empathy makes assessment and interpretation decisions that are perceived by humans to be situationally appropriate. This section reports on an evaluation of CARE models to determine their perceived accuracy. Perceived accuracy tells us whether the behaviors generated by a model are actually perceived to be socially appropriate in practice. Perceived accuracy is an important aspect of empathetic accuracy because, ultimately, we seek to create models of empathy that will generate behaviors that are deemed to be appropriate for a given social context by human observers.

6.1. Method

6.1.1. Participants and Design

In a formal evaluation, thirty-one undergraduate students, in an Institutional Review Board (IRB) of NCSU approved user study, evaluated empathetic responses of the companion agent in video clips from interactions in Treasure Hunt.[3] There were 29 male subjects and 2 female subjects varying in race, ethnicity, and age. 6.5% were aged 18-19, 87.0% were aged 20-24, and 6.5% were aged 25-29.

6.1.2. Materials and Apparatus

For each participant the pre-experiment materials consisted of a consent form, demographic survey, Davis' Interpersonal Reactivity Index questionnaire, Chapin's Social Insight Test (Gough, 1993), and a one-page summary of the construct of empathy. Experiment paper-and-pencil materials consisted of response worksheets for each video clip. The experiment's computerized materials consisted of ten clips of interactions captured from the Treasure Hunt virtual environment. The post-experiment paper-and-pencil materials consisted of a general survey about the participant's experience and opinions on affect in interactive applications, such as games. The demographic survey collected basic information such as gender, age group, ethnicity, marital status, and number of children. Participants completed Davis's Interpersonal Reactivity Index (IRI) to obtain a measure of their empathy (Davis, 1983). Chapin's Social Insight Test quantifies a person's ability to appraise another person by assessing her ability to predict future events involving the other person in interpersonal and social situations. Chapin's Social Insight Test asks subjects to assess twenty-five social dilemmas by selecting the best resolution from the four presented possibilities (Gough 1993). The background document provided the definitions and explanations of empathy from Davis (1994) and Hoffman (2000). Each response worksheet asked subjects the same series of questions about each video clip. Using the response worksheets, subjects evaluated the appropriateness and accuracy of the empathetic emotion, behavior, and timing viewed in

[3] There was no overlap in the 31 participants in the predictive accuracy evaluation (Section 5) with the 31 subjects participating in the perceived accuracy evaluation (Section 6).

the clip, and identified a more appropriate empathetic response, if the participant felt one was applicable.

Each video clip depicted a companion agent exhibiting an empathetic behavior in response to a situation involving another character in the Treasure Island environment. Three types of behaviors were depicted:

- *CARE-generated behaviors*: One set of video clips depicted empathetic responses that were exhibited by companion agents with CARE-induced decision tree models of empathy.
- *Inverse empathetic behaviors*: One set of video clips depicted empathetic responses that were, in effect, the opposite of what CARE recommended. These were determined by identifying the valence of the CARE-generated behavior and then selecting an "opposing" behavior from the classic two-dimensional affective space (Lang, 1995) that had an opposing valence. The inverse empathetic behavior for *excited* was *sad*, the inverse empathetic behavior for *frustrated* was *relaxed*, and the inverse empathetic behavior for *joyful* was *fearful*.
- *Human-generated behaviors*: One set of video clips depicted captures of empathizer-target trainer interactions following the procedure discussed in Section 5, i.e., the behaviors were in fact produced by humans (training empathizers) with the empathizer controls.

Video clips averaged approximately 90 seconds. So that viewers could assess the social context in which an empathetic behavior played out, each clip included several events in the virtual environment leading up to the empathetic behavior, as well as the empathetic behavior itself.

6.2. Procedure

Participants entered a conference room where they were first presented the details of the study and a consent form. They then completed the demographic survey, Davis's IRI questionnaire, and Chapin's Social Insight questionnaire. Next, they read the background on empathy and task directions. Research assistants then fielded any questions from participants regarding empathy and their prescribed task. Participants were then presented, in random order, a series of ten video clips of captured user-interactions in the Treasure Hunt virtual world. There were four clips of CARE-generated behaviors, three clips of inverse empathetic behaviors, and three clips of human-generated behaviors. After viewing each clip, participants completed the associated response worksheet at their own pace. Following the completion of reviewing and responding to all of the video clips, participants completed the post-experiment survey before the study session concluded.

6.3. Results

This section analyzes the study participants' assessments of the empathetic response clips. A variety of ANOVA statistics are presented for results that are statistically significant. Because the tests reported here were performed on discrete data, we report Chi-square test statistics (χ^2), including both likelihood ratio Chi-square and the Pearson

Chi-square values. Fisher's Exact Test is used to find significant p-values at the 95% confidence level ($p < .05$).

The IRI results of participating subjects are reported in Table 1. Participants averaged 16.55 (SD = 4.33) on the Social Insight Test. Gough (1993) reported the results of a study conducted with a similar population consisting of undergraduate engineering students, whom averaged 25.01 (SD = 4.83) on the Social Insight Test. The difference between the subjects in the two studies is not significant ($p < 0.5$).

Table 1. Interpersonal Reactivity Index Results.

Scale	**Mean**	**SD**	**Median**	**Mode**	**Min**	**Max**
Fantasy	16.48	4.75	17	11	8	26
Perspective Taking	17.26	4.77	17	20	6	25
Empathetic Concern	17.87	4.42	18	18	7	28
Personal Distress	9.13	5.19	9	11	0	20
Total	60.74	12.08	62	63	34	84

Analysis of participant responses to video clips depicting CARE-generated behaviors yielded 90.3% of participant responses who agreed that the displayed empathetic emotion was appropriate for the situation, 10.8% agreed clips of inverse empathetic behaviors were appropriate, and 87.1% agreed clips of human-generated behaviors from training episodes were appropriate (Table 2a). Participants also assessed whether the displayed empathetic emotion was the best emotion for the situation. Three-fourths (75.8%) of the participants agreed the displayed empathetic emotion was the best emotion in clips of

Table 2. Analysis of Appropriateness Responses. Grayed-cells indicate no significance ($p < .05$).

	Clip Comparison	**Likelihood Ratio (χ^2)**	**Pearson (χ^2)**
(a)	CARE vs. Inverse	155.13	136.70
	CARE vs. Human	0.56	0.56
	Human vs. Inverse	122.76	108.46
(b)	CARE vs. Inverse	99.75	90.12
	CARE vs. Human	0.20	0.20
	Human vs. Inverse	81.24	74.28
(c)	CARE vs. Inverse	62.51	60.16
	CARE vs. Human	1.12	1.13
	Human vs. Inverse	41.46	39.59

CARE-generated behaviors, while 10.8% agreed for clips of inverse empathetic behaviors, and 73.1% agreed for clips of human-generated behaviors from training episodes (Table 2b). The third response had participants assess the timing of the empathetic behavior (i.e., was there a more appropriate instance in which the companion agent should have been empathetic, or not). Fully 87.9% agreed that the timing of the behavior was appropriate in clips of CARE-generated behaviors, while 37.6% agreed for clips of inverse empathetic behaviors, and 82.8% agreed for clips of human-generated behaviors from training

episodes (Table 2c). Table 2 reports the significant distinctions that can be made between these categories of empathetic behavior clips for each of the above participant responses.

Participants also assessed the accuracy of the displayed empathetic behaviors, using a Likert scale from 0 to 4, with respect to the accuracy of the emotion (Table 3a), the timing (Table 3b), and the overall response (Table 3c). Table 3 reports the results of the participants' accuracy assessment.

Table 3. Empathetic behavior accuracy assessment.

	CARE		**Inverse**		**Human**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
(a)	2.98	1.02	0.77	1.08	2.76	0.97
(b)	3.06	1.00	1.60	1.38	2.76	1.16
(c)	3.04	0.93	0.84	1.09	2.80	0.98

These results suggest that participants perceived the empathetic behaviors controlled by CARE-induced empathy models as being as appropriate and as accurate as human empathizers were in similar situations.

6.4. Discussion

Participant responses to clips of CARE-generated behaviors cannot be statistically distinguished from the responses to clips of human-generated behaviors from training episodes. This result indicates that CARE models generate empathetic behaviors that are similar to those made by humans and are perceived to be situationally appropriate. The fact that participants were able to distinguish, with statistical significance, inverse empathetic behaviors from both CARE-generated behaviors and human-generated behaviors suggests that both CARE models and human models of empathy differ fundamentally from "inverse" empathetic models.

There was no significant effect of psychological instruments (IRI and Social Insight) on participant responses. While a larger study may reveal significant results, it may be the case that measures of one's own empathy does not correspond directly to one's ability to interpret another's empathetic accuracy. A more diverse population that includes subjects beyond undergraduate engineering students may yield different results.

In post-interviews 90.1% of participants indicated that emotions play a valuable role in interactive systems, and 80.1% responded that they would like such systems to account for their own feelings. Most (80.1%) participants indicated that the six emotions displayed in the empathetic behavior clips need to be extended to incorporate additional emotions. *Anger*, *disappointment* (as distinct from *frustration*), and *apathy* were the emotions most frequently suggested for addition to the current model. *Relaxed* was the emotion most frequently suggested for removal from the current model.

7. Conclusions and Future Work

Recent advances in affective reasoning have demonstrated that emotion plays a central role in human cognition and should therefore play an equally important role in synthetic agents. A key affective capability of human social intelligence is empathy. Because empathy is paramount in successful human-human interactions, it may be useful to endow companion agents with the ability to empathize. Empathy modeling requires accurately assessing a social situation context in order to determine (1) if an empathetic reaction is warranted, and (2) if so, what sort of empathetic behavior should be performed.

This article presents a data-driven approach to learning empirically grounded models of empathy from observations of human-human social interactions. In this approach, training data is first generated as a by-product of trainers' interactions in a virtual environment, and models of empathy are induced from the resulting datasets. Critically, the training data employs only observable features, i.e., features that can be directly observed in the environment, so that at runtime, the same features can be used by the empathy models to drive the behavior of companion agents interacting with users. Two complementary types of evaluations, one investigating predictive accuracy and one investigating perceived accuracy, have been conducted on an implemented data-driven empathy modeler. The studies suggest that the data-driven approach offers a promising technique for extending the affective capabilities of synthetic agents to emulate observed human empathetic behavior. Coupling models of social constructs with expressive controls of agent behavior could perhaps contribute to a new generation of socially and emotionally intelligent synthetic agents in the coming years.

In the future, it will be important to investigate mechanisms for varying empathetic responses in a manner that is most appropriate for individual users, perhaps integrating them with tools such as socio-psychologically validated empathy response instruments. It will also be important to devise integrated methods for employing user physiological responses, such as eye gaze tracking, facial feature tracking, posture monitoring, heart rate, galvanic skin response and temperature monitoring (Picard et al., 2001), with empirically grounded models of empathy to further extend their range and increase their accuracy. Conducting large-scale evaluations with a more diverse population is another important direction for future work. Studies with subjects not limited to undergraduate engineering students could more accurately account for a broader range of empathetic abilities. Finally, exploring models of empathy induced from attributes monitored at varying levels of abstraction may yield models that are transferable between different environments.

Acknowledgements

The authors thank Bradford Mott, Seung Lee, and Sunyoung Lee for their contributions to the implementation and evaluation of CARE. The authors also wish to thank Valve Software for authorizing the use of their Source™ engine and SDK.

References

André, E., and Müller, M., 2003. Learning affective behavior. Proceedings of the 10th International Conference on Human-Computer Interaction, Heraklion, Crete, Greece. Lawrence Erlbaum, Mahwah, NJ, 512-516.

André, E., Rist, T., van Mulken, S., Klesen, M., and Blades, S., 2000. The automated design of believable dialogues for animated presentation teams. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., (Eds.), Embodied Conversational Agents, 220-255, MIT Press, Cambridge, MA.

Bates, J., 1994. The role of emotion in believable agents. Technical report CMU-CS-94-136, Carnegie Mellon University, Pittsburgh, PA.

Bates, J., Loyall, B., Reilly, S., 1992. An architecture for action, emotion, and social behavior. Technical report CMU-CS-92-142, Carnegie Mellon University, Pittsburgh, PA.

Baylor, A., 2005. The impact of pedagogical agent image on affective outcomes. Proceedings of the workshop on affective interactions: computers in the loop, International Conference on Intelligent User Interfaces, San Diego, CA.

Bickmore, T., 2003. Relational agents: effecting change through human-computer relationships. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Brave, S., Nass, C., and Hutchinson, K., 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. International Journal of Human-Computer Studies, 62:161-178.

Burleson, W., and Picard, R., 2004. Affective agents: sustaining motivation to learn through failure and a state of stuck. Proceedings of workshop of social and emotional intelligence in learning environments, in conjunction with the 7th International Conference on Intelligent Tutoring Systems, Maceio, Alagoas, Brazil.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., and Yan, H., 2000. Human conversation as a system framework: embodied conversational agents. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., (Eds.), Embodied Conversational Agents, 29-63. MIT Press, Cambridge, MA.

Cavazza, M., Charles, F., and Mead, S., 2002. Interacting with virtual characters in interactive storytelling. Proceedings of the 1st International Conference on Autonomous Agents and Multi-Agent Systems, Bolonga, Italy, ACM Press, 318-325.

Conati, C., 2002. Probabilistic assessment of user's emotions in educational games. Applied Artificial Intelligence, 16:555-575.

Davis, M., 1983. Measuring individual differences in empathy: evidence for a multidimensional approach. Journal of Personality and Social Psychology, 44, 113-126.

Davis, M., 1994. Empathy: A Social Psychological Approach. Brown and Benchmark Publishers, Madison, WI.

De Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., and De Carolis, B., 2003. From Greta's mind to her face: modeling the dynamics of affective states in a conversational embodied agent. International Journal of Human-Computer Studies 59, 81-118.

Elliott, C., 1992. The affective reasoner: a process model of emotions in a multi-agent system. Ph.D. thesis, Northwestern University, Chicago, IL.

Elliott, C., Rickel, J., and Lester, J., 1999. Lifelike pedagogical agents and affective computing: an exploratory synthesis. Artificial Intelligence Today, Lecture Notes in Artificial Intelligence (Sub series of LNCS), Special Volume 1600, Wooldridge, M. and Veloso, M. (eds.), Springer-Verlag, Berlin, 195-212.

Gough, H., 1993. Chapin Social Insight Test Manual. Palo Alto, CA: Consulting Psychologists Press.

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

Hanley, J., and McNeil, B., 1982. The meaning and use of the area under the receiver operating characteristic (roc) curve. Radiology 143:29-36.

Hoffman, M., 2000. Empathy and Moral Development: Implications for Caring and Justice. Cambridge University Press, Cambridge, UK.

Hudlicka, E., 2003. To feel or not to feel: the role of affect in human-computer interaction. International Journal of Human-Computer Studies 59, 1-32.

Ickes, W., 1997. Empathic Accuracy. Guilford Press, New York, NY.

Johnson, L., and Rickel, J., 1998. Steve: an animated pedagogical agent for procedural training in virtual environments. SIGART Bulletin 8:16-21.

Johnson, L., and Rizzo, P., 2004. Politeness in tutoring dialogs: "run the factory, that's what I'd do". Proceedings of the 7th International Conference on Intelligent Tutoring Systems, Maceio, Alagoas, Brazil. Springer-Verlag, 67-76.

Lang, P., 1995. The emotion probe: studies of motivation and attention. American Psychologist, 50(5):372-285, 1995.

Lazarus, R., 1991. Emotion and Adaptation. Oxford University Press, UK.

Lester, J., Towns, S., Callaway, C., Voerman, J., and FitzGerald, P., 2000. Deictic and emotive communication in animated pedagogical agents. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., (Eds.), Embodied Conversational Agents (Eds.), 123-154, MIT Press, Cambridge, MA.

Marsella, S., and Gratch, J., 2003. Modeling coping behavior in virtual humans: don't worry, be happy. Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems (Melbourne, Australia). ACM Press, 313-320.

Nass, C., Isbister, K., and Lee, E., 2000. Truth is beauty: Researching conversational agents. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., (Eds.), Embodied Conversational Agents, 374-402, MIT Press, Cambridge, MA.

Ortony, A., Clore, G., and Collins, A., 1988. The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, UK.

Paiva, A., Dias, J., Sobral, D., Aylett, R., Woods, S., Hall, L., and Zoll, C., 2005. Learning by feeling: evoking empathy with synthetic characters. Applied Artificial Intelligence, 19:235-266.

Picard, R., 1997. Affective Computing. MIT Press, Cambridge, MA.

Porayska-Pomsta, K., and Pain, H., 2004. Providing cognitive and affective scaffolding through teaching strategies: applying linguistic politeness to the educational context. Proceedings of the 7th International Conference on Intelligent Tutoring Systems, Maceio, Alagoas, Brazil. Springer-Verlag, 77-86.

Prendinger, H., and Ishizuka, M., 2005. The empathic companion: a character-based interface that addresses users' affective states. Applied Artificial Intelligence. 19:267-285.

Rickel, J., and Johnson, L., 2000. Task-oriented collaboration with embodied agents in virtual worlds. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., (Eds.), Embodied Conversational Agents, 95-122, MIT Press, Cambridge, MA.

Swartout, W., Gratch, J., Hill Jr., R., Hovy, E., Lindheim, R., Marsella, S., Rickel, J., and Traum, D., 2004. Simulation meets Hollywood: integrating graphics, sound, story and character for immersive simulation. In Stock, O., and Zancznaro, M. (eds.), Multimodal Intelligent Information Presentation, Kluwer.

Witten, I., and Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufman, San Francisco, CA.