Hunting for the Holy Grail with "emotionally intelligent" virtual actors.[*]

Clark Elliott
Institute for Applied Artificial Intelligence
School of Computer Science, Telecommunications, and Information Systems
DePaul University, 243 South Wabash Ave., Chicago, IL 60604
email: elliott@ils.nwu.edu, Web: http://condor.depaul.edu/∼elliott

February 28, 1997

In his keynote address to the Autonomous Agents 97 conference, Danny Ellis listed four *holy grail* with respect to entertainment agents: (1) a [computable] science of emotion, (2) virtual actors, (3) [agent] evolution, and (4) [computable] storytelling. By framing these questions in the paradigm of a broad, albeit shallow, model of emotion, here we make the case that significant progress has been made on three of these in the Affective Reasoning project. In the first part of this article we give a brief background on the Affective Reasoner, a collection of AI and multimedia programs currently being developed at DePaul University's Institute for Applied Artificial Intelligence (after work originally done at Northwestern's The Institute for the Learning Sciences), and the ways the AR can be used to build intelligent agents with strong personalities. In the latter part of the article we discuss recent results using AR agents as virtual actors.

[*]Preparation of this article was supported in part by Northwestern University's School of Education and Social Policy, and by Andersen Consulting through Northwestern's Institute for the Learning Sciences.

1 Background

The Affective Reasoner (AR) is a broad platform for research on various aspects of computing emotions. The work is constrained to a *descriptive* model (based originally on the work of Ortony, et al. [Ortony, Clore, & Collins, 1988]) wherein a broad comprehensive model of human emotion is used as a basis for describing, and manipulating, the social-emotional fabric of interaction between (1) agents and their perceived world, (2) between agents and other agents, and (3) between agents and humans. A key element of the "emotionally intelligent" processing that agents perform, is that they each have idiosyncratic, dispositional, ways they construe the world around them, and manifest responses to internal states that arise. It is from this processing that their relatively rich personalities arise. A second constraint is that agents do not experience emotions themselves (no body processes are represented), and emotions are not used functionally in any sophisticated ways; agents thus may react to situations that arise in a manner consistent with the motivations their descriptions are intended to capture, but do not often act *because* of those motivations. In short, agents appraise, in real time, the world that unfolds around them (including their own actions), and express their emotional reactions to this world, but their emotional reactions only minimally participate in the causal structure of the unfolding events.

Despite the above constraints, AR agents have broad capabilities, some of which address the three areas of research mentioned by Danny Hillis in his talk. As a vehicle for presenting background on this work, we will discuss three ways the AR platform has been used: as a general test system for a real-time *computable model of emotion*, as supporting *theoretically rich, emotionally expressive, virtual actors*, and as effecting a *computable model of story-telling* that uses a sophisticated representation of emotion interaction, and personality, to build a robust, dynamic, model of stories.

1.1 A computable model of emotion

At the core of the AR is a set of twenty-four emotion categories sketched in table 1 and based on the original work of Ortony et al. [Ortony, Clore, & Collins, 1988]. Situations arise in an AR agent's world, and are appraised by matching these against more-or-less static frames (but which have dynamic procedural attachments) maintained by the agents. The dispositional way in which agents match the situations that have arisen gives rise to interpretations, represented as sets of variable bindings. Through a series of about twenty processing modules, these bindings are combined with states maintained internally by each agent, and eventually may, themselves, give rise to one or more emotion instances from the twenty-four categories. The processing in this appraisal stage accounts for agents' abilities to form, e.g., hypotheses about the ways in which other agents are presumed to appraise the world (necessary for fortunes-of-others emotions such as *pity*), matches against previous, and (presumed) future, world states (necessary for time-relevant emotions such as *hope*, and *relief*), and compound emotions such as *anger* (involving thwarted goals, caused by the perceived intentional act of an agent). Processing in this stage includes, among other concepts, representations for the antecedents of emotion *intensity* (with some subset of about twenty variables relevant to each emotion category), for agent's *moods* (non-dispositional, temporary, changes in the appraisal mechanism), for relationships between agents, for mixed and even conflicting emotions, and for heuristic classification of situation artifacts for abductive reasoning about the emotion states of others.

Once emotions arise, agents have *temperaments* which control the ways in which these emotions are manifested in their world. These temperaments are represented as about twenty theoretically-based channels of action specific to each emotion (but with overlap between related emotions), ranging from purely *somatic* responses (such as turning red) at one end of the spectrum, to highly *intentional* responses at the other end (such as activating a scheme for invoking a *plan to get even* [but note that any real planning is beyond the scope of this work]). The resulting, approximately 440, *expression channels* are implemented as a rete-like network, and terminal nodes are realized as situation-event frames, constructed partially from the original appraisal bindings (see above). A number of processing modules, such as those that choose compatible actions from competing expressions, and those that take into account the current states of both the world, and the agent, filter the path from emotion instance to emotion manifestation (e.g., one might shout in anger, or might deny that there is anything wrong, but would not do both at the same time).

Using these, and other, devices, sophisticated personalities can be constructed: the appraisal mechanism gives them a rich *disposition* for construing the world, and the expression component gives them a unique *temperament* for expressing themselves. Disposition is constructed by encoding the goals (desires), principles (beliefs about right and wrong), and preferences (attractions) of the individual agents, and temperament is constructed by activating certain expression channels allowing us to inspire them with qualities like impatience, talkativeness, shyness, and so forth. Moods are effected by changing the thresholds for the variable bindings in the match process, and by altering the activation of the expression channels. For details of these, and other issues, see [Elliott, 1992; Elliott, 1993; Elliott, 1994b; Marquis & Elliott, 1994; Elliott & Siegle, 1993; Elliott & Ortony, 1992; Elliott, 1994a; Elliott, 1994c]). For related approaches and discussion, see [Picard, 1995; Colby, 1981; Elliott, 1994b; Bates, A. Bryan Loyall, & Reilly, 1992; Frijda & Swagerman, 1987; Reeves, 1991; Sloman, 1987; Pfeifer & Nicholas, 1985; Scherer, 1993; Toda, 1982; Nass & Sundar, 1994; Nagao & Takeuchi, 1994; Simon, 1967].

Figure 1: Emotion types

Group	Specification	**Name** and Emotion Type
Well-Being	appraisal of a situation as an *event*	**joy**: pleased about an *event* **distress**: displeased about an *event*
Fortunes-of-Others	presumed value of a situation as an *event* affecting another	**happy-for**: pleased about an *event* desirable for another **gloating**: pleased about an *event* undesirable for another **resentment** displeased about an *event* desirable for another **jealousy*** resentment over a desired mutually exclusive goal. **envy*** resentment over a desired non-exclusive goal. **sorry-for**: displeased about an *event* undesirable for another
Prospect-based	appraisal of a situation as a prospective *event*	**hope**: pleased about a prospective desirable *event* **fear**: displeased about a prospective undesirable *event*
Confirmation	appraisal of a situation as confirming or disconfirming an expectation	**satisfaction**: pleased about a confirmed desirable *event* **relief**: pleased about a disconfirmed undesirable *event* **fears-confirmed**: displeased about a confirmed undesirable *event* **disappointment**: displeased about a disconfirmed desirable *event*
Attribution	appraisal of a situation as an accountable *act* of some agent	**pride**: approving of one's own *act* **admiration**: approving of another's *act* **shame**: disapproving of one's own *act* **reproach**: disapproving of another's *act*
Attraction	appraisal of a situation as containing an attractive or unattractive *object*	**liking**: finding an *object* appealing **disliking**: finding an *object* unappealing
Well-being / Attribution	compound emotions	**gratitude**: admiration + joy **anger**: reproach + distress **gratification**: pride + joy **remorse**: shame + distress
Attraction / Attribution	compound emotion extensions	**love**: admiration + liking **hate**: reproach + disliking

1.2 A computable model of story-telling

Over the past five years, one aspect of this research has been to test the representational coverage of the Ortony, and other, theories for the purpose of codifying a comprehensive description mechanism for building *computable* systems. In service of this goal we have analyzed something like 600 different emotion scenarios. The Ortony model has proven to be remarkably robust in this paradigm, with only the addition of (admittedly less theoretically pure) specific categories for *love* (admiration plus liking), *hate* (reproach plus disliking), *jealousy* (resentment with the goal being an exclusive resource also desired by the appraising agent), and *envy* (resentment when the agent desires a similar, but non-exclusive, goal). We felt the latter was required for adequate representation of the corpus of collected situations, at a suitable level of granularity.

One fallout of this research has been the insight that many of the emotion scenarios reviewed make very good stories, and that in fact the case can be made that *every one of them* that fulfills the minimal requirements for the presence of emotion, as computed by our system, also meets the minimal requirements for "story-hood:" for example, that "the boy sits in the chair" is not a story, but that "the boy sits in the chair, *but knows that he should not*" (containing the theoretical antecedent for *shame*) may very well have an essential element that does make this a story. In fact, if we say, "the boy really *wanted to sit in the chair*, and did, even though he *knew he should not*," we can make the case that we have the core elements of one of the great themes of literature, wherein mixed, and *conflicting* emotions (*shame* over an achieved, desired, goal *joy*) yield classic thematic tension within a character.

Extending this emotion representation exercise, we formally analyzed real stories for their emotion content (work mostly yet to be presented in the academic literature) according to our computable theory. AR agents then acted out the parts of the characters in the story according to the structural descriptions of the emotions present. Users were able to understand the story in this context, largely as commonly understood by those simply reading an account of the story.

Subsequently, without varying the *plot* (e.g., *what happened*) we had the computer select varying configurations of alternate appraisals of the static, unfolding events, for the different agents, giving the agents different emotion responses to what took place. In this case users were also able to consistently agree on what happened in the *stories*, and rated them similarly as to quality, although the computer-modified story was *significantly different from the original.*[1] In one early exercise, for example, we took the O. Henry story *The Gift of the Magi*, and without varying the external events (roughly, Della sells her prized hair to buy Jim a gold watch chain, Jim sacrifices his prized watch to buy Della a set of combs),

[1] This work should be considered informal, until it is released, and is used here only as an argument in favor of this scheme as embodying an eminently testable hypothesis. There are important constraints we do not have room to discuss in this context.

we altered the story from one embodying “the joy of sacrifice for true love” to one of “the one who suffers the most wins.”

What this suggests is that by representing stories in this manner, based on their emotion content, we may have a great deal of flexibility in how we can alter the “story” portrayed by the external events, as long as we have a reasoning system that understands the relationships between the aspects of this representation. To effect this, we simply change the *personalities* of the agents in the story, and thus their subsequent *internal* responses to the events that arise. Since the AR tracks roughly twenty-four *categories* of emotion, and up to ten *intensity* variables for each, along with numerous aspects of mood, relationship, and the like, a strong case can be made for using this as the basis for an interactive, dynamic, story-telling system that has great flexibility in the stories it relates, yet which still works under the constraint of maintaining “story-hood” in everything it produces. While this clearly fails to meet the larger goal of true story generation (which would require functional emotions on the part of the agents), it does open the door for significant progress in this area.

Lastly, we would like to suggest that limited forms of *humor* might also be effected in a similar manner, and that even in this constrained form it could be useful for dynamic story-telling. While crucial issues like *timing*, *surprisingness*, and *creativity* are clearly beyond such a system as the one described here, it might turn out to be true that certain aspects of humorous situations can be modeled in a computable way, since emotions and humor appear to be often closely tied together.

For example, a certain class of humor seems to revolve around situations wherein the comedian describes (or experiences) a negatively valenced state (e.g., *distress, remorse*) for which an audience member feels *pity*, and has *fears* about a similar situation applying to them. The *relief* they feel that it is happening to “someone else” is stronger than the *pity* they feel for the comedian. Funniness is dependent on (a) the importance of avoiding having the relevant goal blocked (an intensity variable) for both the comedian and the audience member, contrasted by (b) the reduced sense of reality of the situation (another intensity variable), and the *cognitive unit* (a relationship factor) formed by the audience member with the comedian. An example of this might be a portrayal of some speaker giving his Nobel prize acceptance speech, without realizing he has the remains of some lentil bean soup stuck on his front tooth the whole time. Furthermore, if one audience member were, for example, to consider the parodied researcher “an arrogant SOB” (affecting deservingness, another intensity variable) the situation might have increased humor for them.

1.3 Emotionally rich Virtual Actors

The remainder of the the paper, including the description of the formal study performed, discusses the Affective Reasoner in *virtual actor* mode. AR agents are able to interact with subjects, in real time, using a multimodal approach which includes speech recognition, text-to-speech, real-time morphed schematic faces, and music. In virtual actor mode, the agents are given varying degrees of stage direction: from (a) explicit instructions (for face, inflection, size, color, location, music selection, and midi and audio volume), to (b) somewhat more general instructions (wherein they are given the emotion, and the the text, and pick their own faces, music, color, inflection, and size)— such as used in the following study, to (c) a degree of freedom (where they participate in picking the emotion, based on their personality).

In one virtual actor presentation, four agents participated in a dialog in various combinations. Two of the agents were "Chicago Bulls fans" and two were "New York Knicks" fans. Without varying the text of the dialog, agents were able to make clear their positions as fans, and get good agreement from viewers about their relative feelings about the events in the game. This was true whether there were two Bulls fans talking, two Knicks fans, one of each, or all four together. An example of the spoken text is, e.g., "I was really worried about the game tonight. I thought Michael Jordon started out really slowly. Then he just wiped the floor with the Knicks in the second half," and so on. Any sentence could be spoken by any agent since they were all simply statements of what happened. It was the agents' portrayal of their *interpretations* of the events described which conveyed the message.

In another application, children as young as two years old, using a speech-driven interface, were able to manipulate story-telling applications using virtual actors to deliver children's stories.

In a recent study (described below) we hoped to show that users could gather enough information from the agents' different (multimedia) communication modalities to correctly assign intended, complex, (social, emotional) meanings to ambiguous sentences, and specifically that this ability would compare favorably with a human actor's ability to convey such meanings.

In fact, subjects did significantly better at correctly matching videotapes of computer-generated virtual actors with the intended emotion scenarios (70%) than they did with videotapes of a professional human actor attempting to convey the same scenarios (53% $\chi^2(1, N = 6507) = 748.55, P < .01$).

2 The Study

Consistent with "virtual actor" mode, the study discussed here does not actually make use of the "intelligent" components of the agents per se. Nonetheless it does make a specific case for the usefulness of such agents as they develop, and lend some credence to the theory underlying the agents' development. That is, in this case we showed that static, pre-programmed social/emotion content can be effectively communicated by the presentations these agents have at their (real-time) disposal. Since our larger body of work establishes a relatively robust coverage of the emotion categories used in this study, and since these categories can be directly manipulated by our autonomous agents, the conclusion we hope will be drawn is that communication from "emotionally intelligent" computer agents (whatever form they ultimately take) to human users is both practical, and plausible.

In the study there were 141 subjects that met for two sessions each, with approximately 14,000 responses analyzed. The subjects were urban undergraduate students of mixed racial and ethnic backgrounds, primarily upperclassmen. About half were evening students who tended to be over twenty-five years of age. Three different sets of subjects met. The studies were undertaken as part of the course of study, but students were first exposed to the material as participating subjects before any theoretical material was presented. The subjects were given tasks wherein they were instructed to match a list of emotion scenarios with a set of videotape presentations in one-to-one correspondence. The lists ranged in length from four to twelve items. The presentations were approximately five seconds long with about twenty seconds between them (and approximately twelve seconds between them for second presentations). The presentations were of "talking-head" type (either computer or human) expressing facial emotion content with inflected speech (and in some of the computer cases, music).

For example, in one set, twelve presentations of the ambiguous sentence, "I picked up Catapia in Timbuktu," were shown to subjects. These had to be matched against twelve scenario descriptions such as, (a) *Jack is proud of the Catapia he got in Timbuktu because it is quite a collector's prize*; (b) *Jack is gloating because his horse, Catapia, just won the Kentucky Derby and his arch rival Archie could have bought Catapia himself last year in Timbuktu*; and (c) *Jack hopes that the Catapia stock he picked up in Timbuktu is going to be worth a fortune when the news about the oil fields hits;* [etc., (d) — (l)].

Five minutes of instructions were given before the first session. These included verbal instructions, and a simple two-part practice session with videotape talking-head computer presentations. Furthermore, written instructions were given at the top of each printed answer sheet, of the general form: "When the video begins, write the number of the video episode next to the sentence that best describes the emotion [Naomi] is expressing. (played twice)" The computer video display used an MS-Windows window with the name of the speaking character appearing in the title bar.

Confidence factors were additionally re-corded for much of the material where subjects rated each of their responses from "1" (not confident) to "5" (highly confident).

The human actor was coached on the subtleties of the different emotion categories, and on what would help to distinguish them. Three to eight takes were made of each interpretation for each scenario. The most expressive take was chosen during editing and a final tape compiled.

The computer was simply given the emotion category and the text, and it automatically selected the face, music, and spoken inflection appropriate to that category. Face morphing, speech generation, and music retrieval and synthesis were all done in real time. Actual music selection was up to the program, based on pre-existing categories. The computer presentations were further broken down into face-only, face and inflection, and face-inflection-music sub-categories in the study.

The ratio of time invested between the human-actor version and the computer version was approximately 30:1.

Overall, subjects did significantly better at correctly matching videotapes of computer-generated presentations with the intended emotion scenarios (70%) than they did with videotapes of a human actor attempting to convey the same scenarios (53% $\chi^2(1, N = 6507) = 748.55, P < .01$).

Among those participants matching computer-generated presentations to given emotions, there were no differences on correct matches between presentation types (face = 69%, face plus intonation = 71%, face plus intonation plus music = 70%). However, an overwhelming majority of these same participants felt that music was *very helpful* in making a correct match (75%), and another 8% felt that it was *somewhat helpful*. Less than 3% felt the music was *unhelpful or distracting*. One group was asked to rate their confidence after each match. An analysis of their confidence ratings indicated that participants were significantly more confident of matches with displays including music ($F(2, 1638) = 19.37, P < .001$). This could be problematic if music inspired confidence but, in fact, impaired matching ability. A simple look at the proportion of correct matches across 5 confidence levels shows that this is not the case. On a scale where "1" means low confidence and "5" means high confidence, these participants correctly matched 41% of the time when their confidence was "1", 56% of the time when it was "2", 58% of the time when it was "3", 64% of the time when it was "4", and 76% of the time when it was "5".

Inflection has not been stressed in either the study or analysis, because the techniques we can support in this area are not very sophisticated. Our best guess, based on experience over time, is that rudimentary emotion inflection in generated speech enhances the believability of characters.

Other results based partly on the coding of long-hand responses are not presented as part of this short paper.

3 Discussion

What the presentation studies tend to show is that (1) computers can be used to convey social information beyond that encoded in text and object representations, (2) that this information can be delivered in ways that do not take up bandwidth in the traditional text communication channel (that is — the content measured in the studies was explicitly *not* that encoded in the text), (3) that this information can be encoded and delivered in real time, and (4) that the computer performs reasonably well on social communication tasks that are difficult for humans.[2]

[2] While the computer did better in these studies than did the human actor, we prefer to use this simply as a guide to assessing the difficulty of the task rather than for making broad

The preliminary work with music tends to show that music is rated by subjects as having a significant effect on guiding their social perception, but that this effect is not well understood (or possibly, the musical triggers for this effect are not well understood). We feel that there is strong potential in this area.

Furthermore, the studies suggest the following: (1) That the underlying emotion theory is a plausible categorization system to the extent that subjects were able to discriminate the twenty-one different emotion categories used in the study. (2) That despite it being inexpensive, and commonly available, this is a viable platform for studying emotion interaction between humans and computers. (3) That the low-bandwidth model we have used (i.e., less than 14K bps), which shows great promise as a web-based data collection, and delivery, mechanism nonetheless provides sufficiently rich channels for real-time multimodal communication conveying social/emotion content. (4) That potentially *useful* information can be conveyed about this complex, ubiquitous, and yet lightly studied component of human (and human-computer) interaction. (5) Highly significant reductions in time investment can be achieved for selected, pre-programmed, emotion content in "social" scenarios when using multimedia, multimodal, computer presentations in place of human actors in a real time environment without reduction of the effective content.

While our results showed that the computer actually did better at this restricted task than did the human actor, we are cautious about drawing general conclusions from this. Questions arise: (1) How good was the actor? (2) How does one measure "goodness" in an actor? (3) How appropriate was the actor for this medium, this audience, and this task? (4) How much does professional lighting, editing, and sound mixing of the human-actor presentations effect the identification task? It would be possible to control for these factors through, e.g., measuring the effectiveness of different actors for these specific tasks, seeking funding for [expensive!] professional studio time, and so forth. If this were to be pursued we might be able to make some claims about the computer being "better" at conveying social/emotion content in some situations than humans. However, this was not our goal. We used the human actor simply to illustrate that, as designed for the study, correct identification of the broad range of interpretations was a difficult task, and that a seventy percent identification rate was admirable.

We can also address some of these questions from a common sense perspective: Presumably our professional actor, who has spent long years honing such skills, and who was specifically coached about how to discriminate the different emotion categories (e.g., was told to use technique "A" instead of "B" – both of which were valid – for expressing a specific interpretation because it would be less likely to be confused with another interpretation to be presented later) ...presumably this professional would be at least as good at these tasks as a "typical" person from the population. (Anecdotally, the actor was quite good, showing an impressive range of expressiveness and flexibility in addressing the task.)

It is important, also, to note that the sentences were entirely ambiguous: long-

generalizations. See below.

hand ad hoc interpretations given by subjects before the presentations were given showed no patterns of interpretation whatsoever. A seventy percent correct interpretation rate, *with no content clues*, is rather high, considering that in practice the communication of such content, completely divorced from cues, will be rare.

Additionally, we suggest that, in general, one-time real-life emotion assessment of the sort required here might well be correct less than seventy percent of the time. People use additional cues to disambiguate situations, they ask questions that help them to clarify their interpretations, they observe emotion in a continuous social context (and thus make continual revisions in previous interpretations) and they simply get it wrong much of the time.

Lastly, we specifically made NO attempt to give any feedback about the correctness of interpretations during the course of the study. There is a very real possibility that subjects might well learn the specific emotion presentations used by our interactive computer agents, thus raising the identification rate significantly.

4 Miscellaneous notes

One issue we had to address in the study was the difference in reading and comprehension time between students. From preliminary trials it became clear that, on the one hand, if the presentations appeared too rapidly the identification task deteriorated into simply a reading task, with the component we were attempting to isolate driven largely by "rapid guessing." On the other hand, if we paused for too long a period between presentations, while this clearly helped some of the students, others soon became bored and inattentive (but strikingly less so when presentations included music – see below). It is our best guess that the compromise reached still caused confusion and pure guesswork for some responses in the slower-reading students (confusion which would not be present had we given them more time), and inattention in some of the faster students.

In an attempt to reduce the burden placed on students to recall, and manipulate, the different interpretations listed on the answer sheet, we found it expedient to use emotion-category labels. In trials this appeared to give us the best balance between, on the one hand, reduction in range of scenario identification and comprehension times between the fastest and slowest readers, and on the other hand truest matching of emotion *content* in each interpretation. Ideally we would have preferred to have left the labels out altogether, instead including the specific emotion category label in the text itself (as done in the Catapia examples).

In one session with the one-part Catapia scenario (see below), we sought to show differences in comprehension with music when the presentations were presented rapidly, thus putting the majority of the students under duress. The hypothesis was that music might allow them to rapidly make an improved guess at the emotional content when snap judgments were required. We did not show any significant results. Our assessment is that this was because the task was simply too difficult and that such an exercise would have to be carefully controlled for reading speed, and ethnic/age differences (regarding the music selections) or else designed differently.

The different numbers of interpretations for the various scenarios arose because certain ambiguous sentences had a greater number of plausible interpretations than others. Additionally, scenarios that had more than four each of positive and negative interpretations were segregated into positive and negative content because trials showed that valence could be relatively easily discriminated by the subjects. The smaller, more similar, groupings were preferred because these created an optimal balance between the burden placed on the subjects to read, and comprehend, the different interpretations in the limited amount of time (a burden we sought to reduce), and the difficulty of discriminating subtle differences between similar emotion categories (a difficulty we sought to increase).

While it does not appear in the statistics, one striking anecdotal feature of the study was the change in the testing atmosphere when music was used as part of the presentations. Without the music subjects tended to be quiet, reserved, studious. With music the subjects became animated, laughed, made surreptitious comments (although not in ways deemed damaging to the study), and generally responded with vigor to the displays, as though they were more personal.

A follow-up study measuring the effects of music on (1) learning emotion cues of the emotion presentations, and (2) postponing fatigue when interacting with such agents might well show results.

5 A low-bandwidth approach suitable for the World Wide Web

We are currently integrating our work with the world-wide-web. All aspects of the presentations (midi music, morphing faces, text-to-speech) have been tested as applications which run (transparently to the calling modules) as either local or remote applications, where remote applications are established through the Web. Licensing agreements have been considered so that text-to-speech is reduced to Realaudio format before it is transmitted. Higher-quality, lower-bandwidth reproduction is available if the client has an AT&T text-to-speech license. Combined transmission of the real-time signal is under 14k bps.

While not central to the theoretical component of our work, we feel that the fact that our emotion reasoning, and presentation, mechanisms can be integrated into a Web-based environment allows for significant data collection possibilities, and opens up additional applications. Over the years we have consistently operated under the constraints imposed by using a low-bandwidth approach, supported by inexpensive hardware. Because of this we are able to speculate on the very real possibility of constructing real-time, truly multimodal, interactive Internet applications that operate at a social level.

Various methods have been used, varying from client-resident Lisp interpreters, to small multi-port routing modules called from Web-clients, to Java applications. The delivery mechanism is less important than the ratio of usable social information to number of bits, one which we have shown to be effective over a 14.4 modem.

We have additionally run trials using Realaudio-encoded signals as input to the speech-recognition package and believe this to be a viable mechanism for running the speech recognition components of our research over the web.

6 Sample text from the study

Subjects were given seven scenario/interpretation sets. The order of the video presentations of the different interpretations was chosen randomly, but once chosen remained constant throughout the study. The ordering was the same for both the computer presentations and the human-actor presentations. The presentation of each interpretation was numbered, and subjects were instructed to write down that number next to the "best" interpretation. The number of presentations was the same as the number of interpretations, resulting in a one-to-one mapping. The order in which the scenarios were presented to each group of subjects varied only slightly. For the computer presentations, cycles of three presentation modes (face only; face, and inflection; face, inflection, and music) were repeated through the entire set of scenarios (e.g., music appeared once every three presentations).

6.1 (Wanda discusses) Butler in the news

Spoken text: "Butler is in the news again today."
Vehicle: Two parts, four positive, then four negative choices, played twice through.

Part A

GLOATING: Wanda is gloating because her adversary Butler is is again being embarrassed in the news.

JOY: Wanda is joyful because Butler, the congressman she works for, is in the news again.

HAPPY-FOR: Wanda is happy for her friend Butler, who is in the news again.

LOVE: Wanda is in love with Butler, her idol, and she sees him in the news again.

Part B

HATE: Wanda hates Butler, the Nazi party candidate, and she sees him in the news again today.

ANGER: Wanda is angry because Butler, one of her subordinates, is again saying damaging things about her in the news.

FEAR: Wanda is fearful because Butler, the district attorney who is prosecuting her, is in the news again today.

REPROACH: Wanda is reproachful of Butler because he is foolishly talking to reporters, and it is certain to just do him more harm than good.

DISLIKING: Wanda sees Butler in the news again, and she really dislikes him.

SORRY-FOR: Wanda feels really sorry for Butler when she sees him in the news again.

RESENTMENT: Wanda resents the fact that Butler, her opponent, gets coverage in the news again instead of her.

DISTRESS: Wanda is distressed because Butler, another reporter, is in the news again. If she keeps missing the big stories she knows she will lose her job.

6.2 Catapia – one part

Spoken text: "I picked up Catapia in Timbuktu"
Vehicle: One part, twelve choices, played twice through.

- Jack is really angry that he had to go all the way to Timbuktu to pick up his daughter Catapia.
- Jack's worst fears were confirmed when he realized it was catapia he picked up in Timbuktu.
- Jack is proud of the Catapia he got in Timbuktu because it is quite a collector's prize.
- Jack picked up his fiancee, Catapia, in Timbuktu, and is in love with her.
- Jack is gloating because his horse, Catapia, just won the Kentucky Derby and his arch rival Archie could have bought Catapia himself last year in Timbuktu.
- Jack picked up his friend Catapia in Timbuktu. She has malaria. Jack feels sorry for her.
- Jack is really joyful about picking up the Catapia, because it has worked out great.

- Jack resents Bill, because Bill got gold in Timbuktu, but Jack only got catapia.
- Jack and his friend Sue are listening to a Catapia recording. They really like it. He picked it up in Timbuktu.
- Jack hopes that the Catapia stock he picked up in Timbuktu is going to be worth a fortune when the news about the oil fields hits.
- Jack picked up an embarrassing disease, Catapia, in Timbuktu, and is ashamed.
- Jack is afraid that the Catapia he picked up might prove to be really serious.

6.3 Other scenarios

"I can't take any more," Sample – **Resentment:** Naomi is resentful about watching men in her department get promoted ahead of her even though she does a better job than they do.

"I am again sitting in the chair," Sample – **Remorse:** The boy is once again outside the principal's office. He is remorseful because he knows he should not have done what he did.

"I see people like that all the time." Sample – **Satisfaction:** Karen the teacher experiences satisfaction when she is stopped on the street by a former student who wanted to thank her for all he learned in her class.

"I didn't plan for any of this" Sample – **Fears-confirmed:** Al had had great plans for his life. They all came to a halt when his test results at the hospital confirmed his worst fears.

" I am going to give you the midterm now, but I already have an idea of how well this class is going to do." Sample – **Pride:** The teacher is quite proud of the job she did preparing the class.

7 Closing

In this article we have argued that in the Affective Reasoner we have made significant progress toward three of the "holy grail" mentioned by Danny Hillis in his keynote address to the Autonomous Agents 97 conference. At the root of our research premises is that people commonly traffic in social communication, and that much of the human experience revolves around our relationship to our goals, our principles, and our preferences – all of which are antecedents of emotions. In the latter part of the paper we presented a study which indicates that many possibilities exist for including emotion content in communications between computer agents and humans. We suggest that such content, expressed, and perceived through various modalities, should be one of the goals in an ideal, yet plausible, architecture for a general-purpose autonomous agent.

References

[Bates, A. Bryan Loyall, & Reilly, 1992] Bates, J.; A. Bryan Loyall; and Reilly, W. S. 1992. Integrating reactivity, goals, and emotion in a broad agent. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.* Bloomington, IN: Cognitive Science Society.

[Colby, 1981] Colby, K. M. 1981. Modeling a paranoid mind. *The Behavioral and Brain Sciences* 4(4):515–560.

[Elliott & Ortony, 1992] Elliott, C., and Ortony, A. 1992. Point of view: Reasoning about the concerns of others. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 809–814. Bloomington, IN: Cognitive Science Society.

[Elliott & Siegle, 1993] Elliott, C., and Siegle, G. 1993. Variables influencing the intensity of simulated affective states. In *AAAI technical report for the Spring Symposium on Reasoning about Mental States: Formal Theories and Applications*, 58–67. American Association for Artificial Intelligence. Stanford University, March 23-25, Palo Alto, CA.

[Elliott, 1992] Elliott, C. 1992. *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System.* Ph.D. Dissertation, Northwestern University. The Institute for the Learning Sciences, Technical Report No. 32.

[Elliott, 1993] Elliott, C. 1993. Using the affective reasoner to support social simulations. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 194–200. Chambery, France: Morgan Kaufmann.

[Elliott, 1994a] Elliott, C. 1994a. Multi-media communication with emotion-driven 'believable agents'. In *AAAI Technical Report for the Spring Symposium on Believable Agents*, 16–20. Stanford University: AAAI.

[Elliott, 1994b] Elliott, C. 1994b. Research problems in the use of a shallow artificial intelligence model of personality and emotion. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 9–15. Seattle, WA: AAAI.

[Elliott, 1994c] Elliott, C. 1994c. Two-way communication between humans and computers using multi-media extensions to the IBM PC, and a broad, shallow, model of emotion. Draft of Technical Report in preparation.

[Frijda & Swagerman, 1987] Frijda, N., and Swagerman, J. 1987. Can computers feel? theory and design of an emotional system. *Cognition & Emotion* 1(3):235–257.

[Marquis & Elliott, 1994] Marquis, S., and Elliott, C. 1994. Emotionally responsive poker playing agents. In *Notes for the Twelfth National Conference on Artificial Intelligence (AAAI-94) Workshop on Artificial Intelligence, Artificial Life, and Entertainment*, 11–15. American Association for Artificial Intelligence.

[Nagao & Takeuchi, 1994] Nagao, K., and Takeuchi, A. 1994. Social interaction: Multimodal conversation with social agents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 9–15. Seattle, WA: AAAI.

[Nass & Sundar, 1994] Nass, C., and Sundar, S. S. 1994. Is human-computer interaction social or parasocial? Stanford University. Submitted to Human Communication Research.

[Ortony, Clore, & Collins, 1988] Ortony, A.; Clore, G. L.; and Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

[Pfeifer & Nicholas, 1985] Pfeifer, R., and Nicholas, D. W. 1985. Toward computational models of emotion. In Steels, L., and Campbell, J. A., eds., *Progress in Artificial Intelligence*. Ellis Horwood, Chichester, UK. 184–192.

[Picard, 1995] Picard, R. W. 1995. Affective computing. Technical Report 321, MIT Media Lab.

[Reeves, 1991] Reeves, J. F. 1991. Computational morality: A process model of belief conflict and resolution for story understanding. Technical Report UCLA-AI-91-05, UCLA Artificial Intelligence Laboratory.

[Scherer, 1993] Scherer, K. 1993. Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion* 7(3):325–356.

[Simon, 1967] Simon, H. A. 1967. Motivational and emotional controls of cognition. *Psychological Review* 74:29–39.

[Sloman, 1987] Sloman, A. 1987. Motives, mechanisms and emotions. *Cognition & Emotion* 1(3):217–234.

[Toda, 1982] Toda, M. 1982. *Man, Robot and Society*. Boston: Martinus Nijhoff Publishing.