DRAFT: Components of two-way emotion communciation between humans and computers using a broad, rudimentary, model of affect and personality.*

Clark Elliott
Institute for Applied Artificial Intelligence
DePaul University
243 South Wabash Avenue
Chicago, IL 60604
and
School of Education and Social Policy
Northwestern University
Evanston, IL
email: elliott@ils.nwu.edu

July 7, 1995

Abstract

The use of computers for translating human affective communication into symbolic form, and for conveying rudimentary simulated emotions, has been little explored. In this short article we introduce an emotion reasoning engine called the Affective Reasoner based, in part, on the ideas of Ortony, et al. [Ortony *et al.*, 1988], along with its recent multi-media extensions, with which we attempt to address this problem. In the paper we suggest that users may best be maximally *engaged* with the computer, for certain tasks, by taking advantage of what appear to be innate human tendencies toward social, and emotional, interchange. We discuss two preliminary areas of exploration, based on speech recognition technology, in which we develop the spoken communication lexicon between user and computer. The first of

*Preparation of this article was supported in part by Andersen Consulting through Northwestern University's Institute for the Learning Sciences. Citation for finished paper (in Japanese) appears in the bibliography. THIS IS ONLY A WORKING COPY.

these has to do with parsing emotion inflection in simple human utterances. The second has to do with interactively extending a base lexicon of spoken phrases that includes 198 emotion words, as well as tokens describing relationship, mood, and emotional intensity, so that users may add simple non-emotion tokens of their own choosing. This allows them to communicate about emotion situations, in diverse domains, without programmer intervention. Lastly we discuss emotionally expressive channels the multi-media computer has at its disposal. Among these are rudimentary emotionally inflected speech; indexed sub-second access to affect-inducing, and enhancing, music; schematic facial expression supporting over 60 expressions, as well as over 3000 dynamically constructed morphs; and explicit emotion content in utterances generated by the underlying emotion engine. The current implementation runs on an IBM (compatible) PC while still maintaining sub-second response time to (spoken) user utterances, to dynamic generation of new content in spoken text, to morphed changes in facial expression, and to retrieval and presentation of affect-bearing music, constraints we consider essential for supporting plausible affective interaction with the user.

1 Introduction

In this article we will argue that existing technology makes it possible to develop limited emotion communication between computers and humans, and that such work is motivated by allowing us to support a greater capacity for *engaging* the user in complex tasks. In support of this argument we will present a brief background sketch on components of a multi-faceted computer program known as the Affective Reasoner (hereafter referred to as the AR), a large Artificial Intelligence LISP program recently ported to run on an IBM (compatible) 66 Mhz multi-media PC. Following this we will discuss aspects of recent multi-media additions to the AR that allow users to both speak to, and listen to, the computer, in marginally inflected speech; to hear music that powerfully reflects coarse-grained emotion states; and to observe affect-bearing, dynamically morphed, schematic facial representations of the emotional states of the computerized agents.

2 Complexity without confusion

One of our motivations in developing this work is that maximally *engaging* the user in certain classes of computer tasks requires establishing a delicate balance between interesting complexity on the one hand, and confusion on the other. In this section, we discuss what appears to be innate human tendencies toward social modes of interaction, even with unsophistated computerized agents, and how such tendencies, and adaptive social skills, may allow us to add desired complexity to computer systems without endangering the user's comprehension of domain tasks. Lastly, we discuss important constraints on such systems.

2.1 Framing the problem

A postulate of our present work is that humans are quite clever when it comes to understanding emotional cues.[1] In our research, we hope to show that interfaces that use these cues in an intelligent and responsible way will allow for a natural interchange of affective information not normally associated with the use of computers. There are many ways such cues can be given. For example, cartoon characters, such as appear in newspaper comics, are representationally very simple, yet convey much to the reader about the social situation, and the affective state, of the cartoon agent. Multi-media computers, as will be discussed, allow for numerous avenues of dynamically constructed, *responsive*, emotion expression. Among these are animated graphics, inflected spoken text, and music.

Another postulate is that humans are *motivated* to communicate emotionally, and to establish some simple rapport, with a computerized agent, so long as they believe that the agent has at least a minimal understanding of what they are trying to express. Eliza-like programs have long been known to marginally engage some users. However, even the minimal impression of understanding does not always seem to be necessary, since humans, being social creatures, appear to have a tendency to anthropomorphize many different types of objects, such as characters in computer games, characters in MUDs, and even non-animate objects to which they must somehow relate. That human-computer interaction seems to be inherently social is empirically supported by the work of Nass, and others, at Stanford[Nass and Sundar, 1994; Nass and Moon, 1994]. It is our position (and our preliminary efforts have been encouraging), that programs that have some *real* understanding of human emotion, even if highly restricted, will be embraced by users, to the extent that this can be supported in a natural way.

In support of this, consider the following anecdote. At the 1994 AAAI conference in Seattle, Katashi Nagao, and Akikazu Takeuchi showed a video clip of two agents having a "conversation" with an animated video face[Nagao and Takeuchi, 1994]. The face turned toward each of the humans when speaking to, or listening to, each of them. While neither orientation morphing of 3D faces, or movement tracking using a single camera, are new technology, what was remarkable was the response of the audience, which can best be described as being *transfixed* by this portion of the presentation.[2] It is our experience that this is a typical response to agents that present human, lifelike, qualities, and not at all typical as a level of attentiveness for conference talks.

It is further our belief that users will be more forgiving of the idiosyncrasies of emotionally intelligent agents than they are of non-animate computerized objects. This is informally supported in our own observations of users with our emotional agents, where, for example, storytelling agents may not correctly follow directions,

[1]Although we envision that platforms such as the one we are developing will allow this to be tested empirically

[2]I called another observer's attention to the fact that virtually every face in the room was glued to the video presentation.

or may have inappropriate "emotions". This is also supported by the observations of Pattie Maes: In discussing her group's interactive animated video system in presentations, at both the AAAI94 Spring Symposium at Stanford, and at AAAI94 in Seattle [Blumberg, 1994; Maes *et al.*, 1994a; Maes *et al.*, 1994b], she reported that users seemed more forgiving of the technical difficulties of communicating with the animated characters (such as when directing a cartoon child to go away) than they did of the difficulties in communicating with the non-animate objects (such as picking up an animated block). This seems to be because people are used to animate beings sometimes following their own, opaque, internal motivations, whereas they expect inanimate objects to always rigidly obey the laws of nature.

To maximise intelligence in a computer *system* (i.e., where the paradigm includes the user as well as the computer) we must make best use of the relative strengths of the different components, and minimize the burden placed on weak areas of the various components. To this end, an effective system attempting affective interchange between user and computer will establish minimal, but consistent, emotion and personality representation in the computer, and place the primary burden of communication on the user. That is, if users are motivated to communicate with the computer, and believe that it has some minimal understanding of the affective domain, they are best able to adapt to, and assign consistent meaning to, a restricted lexicon: humans are good at this, computers are not. The key to success in placing this communication burden on the users is that they be able to adapt to a system that *does not itself intrude into the interactions*, and thus can support natural, albeit, adapted, intercourse. To this end we have had improved success with, and are encouraged by, the recent advances in speech recognition technology.

2.2 Constraints on such a system

There are three major areas in which such systems must be constrained. First, like systems which use a natural-language understanding front end, and automated tutoring systems, they must only very selectively give the user the impression that there is understanding where none exists. To do so otherwise will invariably lead to frustration, and distrust, on the part of the user, and cause a gap in the believability of the automated agent. It should be noted however, that carefully constraining what the user feels the agent knows, and is capable of, is a very different problem from establishing maximal believability through whatever means possible. For example, characters in stories may be seem very *real* without the reader having misconceptions that they are nonetheless simply characters. Often, the fewer the expectations the user has, the more likely it is that believability and engagement can occur.[3]

Second, the emotion machinery must not get in the way of cognitive uses of the computer. To do so would soon result only in annoying the user. Except for

[3]See also the discussion of schematic facial representations below.

systems specifically designed to study affect and personality, emotion and social capabilities must enhance, but not interfere with, the reasons a user is using the computer in the first place. The best user interfaces are seamless. That is, the best computer interfaces are those which allow the user to perform complex tasks with minimal interaction with, or awareness of, the communication channel between the user and the data or knowledge manipulated by the computer. This is particularly important for (emotion and personality) components that, in themselves, are not designed to present precise information, and of computer agents which may respond according to their own concerns as well as the user's.

Lastly, although possibly premature, we must consider the issue of morality. The solicitation of emotion states in users can easily become ethically suspect, and emotion-bearing communications from users, which are easily recorded in a computer system, can be quite personal. How might a user feel if she discovered that she thought she was communicating on an emotional level with a person, but discovered that she was instead communicating with a machine? These issues are not so far off as one might believe: MUD administrators, for example, have reported a high degree of personal and emotion content in interactions with these (social, remote multi-user) games.[4] In addition, with our own system, we have, for example, observed children grow quite disturbed (as one can well imagine) when the computer responds to them with an angry (animated) face, fierce music, an angry voice, and questions about their behavior.

In summary, one of our motivations in developing this work is the belief that there is broad applicability in engaging the user in rich, socially complex, domains. In such domains there is a delicate balance between, on the one hand, losing the user's interest because a system is too simple, and on the other hand, losing the user's interest because the system is too complex. We have postulated that humans have innate social, and emotional, skills which allow us to raise the threshold level of complexity of such systems without overly reducing the user's aliblity to comprehend application domain tasks. We have suggested that humans are well able to understand simple emotional cues, such as cartoon representations of faces, that humans are *motivated* to communicate on a personal and emotional level if they believe the computer can understand such communications, that they will tend to be forgiving of the idiosyncrasies of computerized agents who appear to have their own personalities, and that a human-computer system should take advantage of the relative strengths of each component: computers to manipulate content, and humans to adapt to the computer's mode of communication. Finally we discussed three non-prohibitive constraints concerning the believability, unintrusiveness, and morality of such systems.

[4] Personal communications with four MUD administrators.

3 Background on the Affective Reasoner

3.1 General overview

A central assumption of the system, whose description follows, is that simulations of interactions between people should incorporate a model of affect and personality. Much human interaction revolves around people's individual needs and goals. These lead to idiosyncratic, internally motivated, behavior, and to emotional responses to situations that arise. If these concepts are not captured then we will likely end up simulating agents that are best bland, lifeless, and unrealistic, a problem largely ignored in AI.[5]

The Affective Reasoner (hereafter AR) is a simulation platform that embraces a wide range of emotion-reasoning issues. In the current research, various worlds are simulated, and populated with agents capable of participating in emotional episodes based on their concerns. Agents are given unique dispositions modeled as a hierarchical set of *appraisal frames*. These frames represent their individual goals, principles, preferences, and moods. Combinations of the appraisal frames are used to interpret situations that unfold in the simulation. These interpretations, in turn, are characterized in terms of the way they may or may not meet the eliciting conditions for emotions. In some cases emotions result, which then may be expressed, through a set of approximately 450 selectively activated behavioral channels, in ways that are observable by other agents, and as new simulation events which might perturb future situations. Additionally, agents use a case-based heuristic classification system to reason about the emotions of other agents (including the user), and to build representations of those other agents' personalities that will help them to predict and explain future emotion episodes involving those observed agents. Extensions to the original system were later added to generate different emotion intensities based on three broad categories of approximately twenty variables in the eliciting situations, and to partially order the action-expression paths according to different emotion intensities. Because of space limitations we omit discussions of personality representation, emotion-based action generation, relationships between agents, emotion intensity representation, case-based heuristic classification of the emotions of other agents, affect-based music indexing, and a broad number of exploratory applications in which the AR has been used (but see also [Elliott, 1992; Elliott, 1993; Elliott, 1994c; Marquis and Elliott, 1994; Elliott and Siegle, 1993; Elliott and Ortony, 1992; Elliott, 1994b; Elliott, 1994d]).

[5]But see [Bates *et al.*, 1992; Frijda and Swagerman, 1987; Reeves, 1991; Sloman, 1987; Birnbaum and Collins, 1984; Pfeifer and Nicholas, 1985; Scherer, 1993; Toda, 1982] for interesting approaches to related problems.

3.2 The emotion eliciting condition theory

Embodied in the AR is a set of rules for mapping from emotion eliciting conditions to *emotion types*, based on the work of Ortony, Clore, and Collins [Ortony *et al.*, 1988]. Their theory has been adapted for the simulation, and extended from twenty-two to twenty-four categories of valenced reactions to situations, as outlined in figure 1. Each of the twenty-four emotion types has a set of eliciting conditions. When the eliciting conditions are met, and various thresholds have been crossed, corresponding emotions result. A key element of the theory is that the way an emotion-eliciting situation maps to eliciting conditions depends on how an individual agent interprets that situation. For example, suppose that a judge rules that title to a piece of land belongs to the family that has lived on it for the past two hundred years, rather than the recently discovered descendants of the owner in whose name the land originally was held. One family sees this as a praiseworthy judgment, upholding the principles of common sense, the other family sees this as blameworthy, violating the principles of inheritance. In both cases the act is the same; it is only the *construal* of the situation which is different.

It is important to note that within the underlying emotion theory, the words naming the individual emotion types are merely used for the convenience of the reader, and it should be understood that the emotion types are distinct from any idiosyncratic properties implied by that word. For example, *fear* names the types of emotions described only as *displeased about a prospective undesirable event*, which could be mild agitation as well as stark terror. Within each emotion type, emotion *instances* are tracked, and generated, so that the range of "emotional experience" for the automated agent is quite broad. For example, within the emotion type *joy* it is possible to instantiate the emotion instances representing *delight, joy, elation, bliss, ecstasy, rapture*, and so forth.

4 Developing the communication lexicon.

In the first part of this article we briefly motivated our research, and presented the shallowest of summaries of the Affective Reasoner's infrastructure[6]. In the remaining space we discuss several components of the interface which allows us to communicate, as users, with the emotionally intelligent automated agents.

4.1 Speech recognition

4.1.1 Detecting, and using, minimal emotion cues in spoken dialog with the computer

In recent versions of the Affective Reasoner we do minimal detection of user emotional inflection for limited phrases. The method we use for achieving this is somewhat baroque, but surprisingly consistent. Direct detection of voice inflection is not intended to be supported by the commercial software we are using, and in fact is actually considered a deficiency. (That is, for a *command* interface, the intended primary purpose of the speech recognition software, inflection should be ignored.) To work around this, we added extra "words" to certain base vocabularies such as, for example, "Menu-choice-A-sad" and "Menu-choice-A-angry." We then trained these words to be the spoken words "Menu-choice-A" said in a sad voice, and "Menu-choice-A" said in an angry voice as two independent sound-patterns. After parsing, the emotion-inflection is siphoned off, and both map to "Menu-choice-A." In this way we separate the affect content of utterances from the cognitive content, and can make use of each in appropriate ways.

Though cumbersome, this has allowed us, with a relatively high degree of accuracy (although we have only tested this with a limited set of phrases), to identify the seven emotion categories *anger, hatred, sadness, love, joy, fear, and neutral*, in user utterances. Our preliminary exercises included, among others, speaking the sentence, "Hello Sam, I want to talk to you," (broken up into three tokens) and the menu choices, "*Menu-choice-X*," where X was A through E. In these exercises we did our best to communicate our emotional state to the computer. We were not manipulative, using tricks, but rather slightly theatrical, as though attempting to express human emotion on the stage. The process might best be described as one of speaking to someone who is slightly hard of hearing, but not overly so.

Let us consider now, as an illustration, the affect channels that have been opened from the user to the computer, based only on inflected menu choices, and also what sort of burden this places on the user before it can be used. Five menu choices, each with seven inflections, means that the user must train thirty-five phrases, which takes less than five minutes. Once this is done, the user can use the resultant voice file for *all subsequent applications* without retraining. Nor are

[6] For example, a satisfactory exposition of even the intensity variables is longer than this article in its entirety.

Figure 1: Emotion types (Table adapted from [O'Rorke and Ortony, 1992] and [Elliott, 1992])

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 **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

the menu choices specific to any application: they can be used as well for asking the user to answer questions about military anecdotes as they can for playing computerized poker, or subscribing to computerized math tutoring, as long as the responses are limited to *portions* of the application requiring a choice-driven interface.[7] This also opens up a number of possibilities with respect to automated tutors, such as their relationship to users (e.g., competitive, sympathetic, etc.), and the timing of their intervention, discussion of which is beyond the scope of this paper, but which is addressed elsewhere [Elliott, 1994d].

It is important to note that this is not an attempt to parse affective information that the user is trying to mask (although we see no reason why research could not be developed along these lines). Rather it is an attempt to open channels for the user to communicate affective information (whether intentional, or merely casual) to a computerized partner.

Since our agents also speak in marginally inflected ways, it has been possible, to some degree, to have the agents communicate, and detect emotion inflection, entirely through aural channels, using utterances that are also understandable to the user. We achieve this by simply training the speech recognition package using computerized speech generation. In this way we have had simple dialogs between multiple computer agents, and the user, wherein inflected speech in both the user, and computer agents, is detected by all three (automated, and human) agents.

4.1.2 Extending the Lexicon

The base emotion lexicon recognized as words, and phrases, by the Affective Reasoner was originally drawn from [Ortony *et al.*, 1987] and elsewhere. It allows the AR to aurally parse such sentences as, "I felt great rage when they did it," "I was a bit shaken when she did it," and "I will be somewhat displeased if you do this."[8] This allows the computer to respond intelligently to such utterances by recalling other instances that had similar simple affective structure (e.g., "I believe that you have had a preservation goal blocked. I know of *someone who once had heart trouble* which is also an instance of a preservation goal being blocked. Is this similar to your experience?"), or by responding sympathetically ("I am sorry you are not happy") [Elliott, 1994c].

Assuming that the users wish to communicate with the computer (as discussed above), we are still severely limited by the lexicon allowed by the speech recognition package. The emotion component of the interface is sufficient for a decent amount of emotion understanding, but a typical communication will include real-world content about which the computer has no knowledge. For example, given the utterance "I am angry at my friend Ruth," the computer can recognize the speech tokens for *angry at* and *friend*, but has no knowledge of Ruth. This deficiency

[7]Note that we are not arguing that choice-driven interfaces be encouraged. We are only illustrating how easily affect detection can be added as a *component* of some larger system.

[8]In preliminary exercises were able to correctly parse 188 of 198 phrases containing emotion words, on the first pass.

comes in two parts. First, the speech recognition package has not been trained to recognize the word "Ruth," and second, even if it did recognize the word, it would not know where to place it in the lexicon.

To work around this problem, until a more generalized speech recognition package becomes available, we simply spell new words to the computer. It verifies the word by saying it back to us (sometimes this requires tweaking in the text-to-speech pronunciation library, but can always be ignored in the way one can ignore the pronunciation difficulties of children). Following this, under LISP control, we queue a series of keyboard commands which automatically takes us through a short training session for the word, recompiles the voice recognition package, reloads it and starts it running again.[9] Once this is done, the new word can now be recognized by the computer. In this way the first deficiency has been temporarily addressed.

Following the word recognition phase, the user can be asked a guided series of questions (aurally) allowing the word to be placed correctly in the lexicon. For this we have used a very simple ATN model. Our purpose has been, at this stage, merely to demonstrate that a significant and dynamic emotion lexicon can be established. In practice, the questions can be brief, and since they are asked verbally they appear relatively natural and conversational.

In our example above, the computer knows that the missing word identifies an agent in the *friend* relationship with the user. It also can be relatively confident that *Ruth* is a proper noun. This interchange then only requires a simple verification, "Ruth is the name of your friend, right?" and is made slightly more accessible by an additional gender query, "Is Ruth a female?" (in English, the pronoun "he" is technically allowed to be non-gender-specific, so all agents not known to be female are referred to as "he."). Once this has been established, a series of other interchanges has now been made possible. To wit:

- Clark: "Ruth is unhappy."
- Sam (The computerized agent): "I am sorry to hear that. You must feel sorry for her."

Explanation: Sam (the computer) knows the friendship relationship. Since Ruth is sad, it is plausible that I, her friend, am *sorry-for* her. During a later interaction, perhaps on another day (unsolicited):

- Computer: 'How is Ruth?"
- Clark: "Ruth is jealous of her rival"
- Computer: How much is Ruth feeling this resentment?

[9] The entire process can take less than a minute for several words, which, in practice, is not too cumbersome.

- Clark: Very much.
- Sam: "Perhaps you are worried about Ruth"

Explanation: Sam has stored the fact that Clark is friends with Ruth. In Sam's understanding, jealousy maps to the emotion type, *resentment*. Sam understands that Ruth is in a negatively valenced state, that the intensity is high, and that it is plausible that Clark is worried about Ruth. (Highly negative states can lead to future piggy-back negative states that threaten preservation goals).[10]

Although having to spell words to the computer, and adding tokens to the lexicon one by one, greatly limits communication, this does nonetheless represent steps in the right direction: (a) Aside from the established emotion lexicon, the limited extensions are determined by the user, and can focus on specific areas of interest. This allows the user to teach the computer key words in a new vocabulary relatively quickly. (b) The new vocabulary accumulates, so that once a word has been trained to the system, it continues to be understood, no matter which context is being used. (c) Complexity focuses on the emotion meta-domain, so that a relatively rich affect component can be developed using only a few tokens to identify agents, principles, objects, simple goals, and so forth. (That is, for example, while there are an unlimited number of reasons to be angry at some other agent, only one token is needed to identify that other agent, and there are a limited number of ways to be angry.) (d) even without possession of even the most rudimentary common sense, or reasoning capabilities (e.g., being able to understand that having one's basement flood is likely to be a negative event), the computer still can intelligently process a great deal about directly communicated emotions, such as components of intensity, likely antecedents of the emotion, probable expressions of the emotion, relationships, and so forth.

In addition to the obvious constraints, there are some additional drawbacks that may be less apparent: (a) Speech recognition of individual words is less accurate than when they are included in phrases. To train the words as part of phrases would be burdensome for the user. We consider this to be something that will gradually become less important as speech recognition packages using the new technology improve. (b) The original emotion lexicon, which is quite extensive, has to be trained by each user, once. (c) There is no intelligence for understanding anecdotes the user may wish to relate, only explicitly stated instantiations of emotion instances within them. Neither is there any sort of narrative understanding,

[10]For these exercises our implementation of the cognitive component of the response (e.g., that Sam asks after Ruth, and that Sam asks about the intensity of Ruth's resentment) is relatively ad hoc, and uninteresting. The emotion content of the responses and the relatively deep emotion reasoning about the meaning of the simple responses, is significantly more sophisticated. Our focus has been entirely on the development of emotion representation, and not at all on discourse generation.

so communications about objects or agents appear to the "listening" agent as part of one running, global, information stream. (d) We have not yet formalized any of the ad hoc components of a developing emotion-to-emotion theory, so there is little flexibility in understanding elements of emotion discourse such as, for example, that anger in one agent can often cause anger, or fear, in another agent.

4.2 Multiple-channel expression of the "emotions" of automated agents

Our recent multi-media extensions to the AR allow us to take a multiple-channel approach to presenting "emotion" expression to the user. The idea is that even though each channel in itself is impoverished, together they can give enough clues for the user to make plausible, consistent, interpretations of the agent's internal affective state. Note that this is not a perfect process in the human world either, since we often are mistaken about what others are actually feeling. In our favor is that the automated characters tend to be very consistent in the way they express certain states (although the states themselves vary considerably), and this allows the user to learn their cues, over time.

One advantage of using multiple channels is that they have different strengths, and tend to compliment one another. The following comments on this interrelationship are not based on formal assessments, but rather on observations of a reasonable degree of user interaction with the system. Our goal in presenting these brief reflections is not to report studies on the efficatiousness of these communication channels, but rather to argue that platforms such as this provide ample room for study and experimentation.

Our observations include the following: (a) The schematic facial expressions give the most immediately identifiable cues as to emotion state. Positive and negative valence is immediately apparent. For some emotions, such as *anger* and *fear*, the emotion itself is readily apparent. For other emotions, such as *hope* and *satisfaction*, a large percentage of other emotions are *eliminated* from consideration. The same is true of emotion intensity: for some emotions, such as *joy*, and *distress*, the varying intensities are immediately discernible, for others, especially when the specific emotion itself has not been identified, certain intensities are merely ruled out. (b) A faster speed to the facial morphing tends to enhance the perception that the emotion is more intense. (c) The longer the emotion takes to decay, the greater the tendency to perceive it as having been more intense. (d) When a face grows larger, the expression tends to be perceived as more intense, and vice versa (e) Emotionally inflected speech adds *immediacy* to the identified facial expression. Because of its real-time, ephemeral, qualities, it seems to demand greater short term attention. In addition, the user is forced to be attentive when listening to what the agent is saying. It adds believability to the emotion. (f) Explicit emotional content in the utterances is the most completely comprehensible, of course, but seems to carry little immediacy. That is, when an agent tells us that he is *ecstatic* about the outcome of some event, we might believe him, but we are not very *engaged* by this static information. This can best be illustrated by imagining the content merely presented as text-based information on the screen. (g) Music is perhaps the least specific, but also appears, in many ways, to be the most engaging, the most visceral. To the extent that other cues identify the emotion, emotion-bearing music *enhances its perception*. There will never be agreement that *any* piece of music explicitly expresses a particular emotion for all users, however it is easy to find a *high* degree of agreement that a particular piece does *not* represent a particular group of emotions, thus narrowing the range of possible emotions actually represented, by elimination. Music commonly identified with specific external scenarios (such as for well-known movies), tends to detract from engagement with the computerized character. (h) Increased music volume seems to suggest greater emotional intensity.

In addition, we also have access to the following dimensions of which we have yet to make significant use: (i) Measured, *rate* of increase or decay in the size of the face, over time, (j) the full range of both background and foreground colors for the faces. (k) a wide range of MIDI controller changes with which we can dynamically alter instrumentation, pitch, and so forth, in the music.

Taken as a whole, we find this to be a rich mechanism for giving the user cues to the state of the automated agents. Nor are we limited to presenting affect consistently across channels. In our poker-playing agents, for example, we use different channels, in idiosyncratically combined ways, depending on the agent, to represent cognitive expression on the one hand, and possibly inconsistent emotional expression on the other. For example, an agent who is bluffing might well attempt to project confidence in speech and facial expression, but might simultaneously "subliminally" express fear through changes in facial expression decay, or choice of expressive music.

Also, it should be noted that the *speed* with which each of these channels may be manipulated during interaction with the user is significant: Any of the 3000 or so morphs can be initiated, changed in midstream to a new destination, sped up, slowed down, and so forth, in hundredths of a second. Simultaneously, the mouth can be made to move, and the position of the face, the color of the face, and the size of the face can be altered. Text can be constructed according to the speed of the processor's ability to derive and manipulate strings, and sent to the speech module in about a second. Access time to *any* portion of the (currently approximately 200 hours worth of) midi music is about a second (and significantly less for about four cached concurrent files). When combined with the often sub-second response time of the speech recognition package, we feel a case can be made for using these system features to support research on real-time emotion interaction with the user.

4.3 The schematic facial expression model

Sophisticated, dynamic, three-dimensional models of faces, such as that shown by Pelechaud at IJCAI93, and Nagao and Takeuchi at AAAI94, have the power to transfix audiences and convey expressive information[Pelachaud *et al.*, 1993; Nagao and Takeuchi, 1994]. However, it is also true that many emotion cues can be abstracted and presented in schematic format. Because the computational flexibility of such faces is much greater, and because symbolic intentionality of expression is so clear, we have used this latter mechanism in our own work.

The faces we use are modeled, originally, after the work of an as yet unidentified artist.[11] From many network queries, and phone conversations, however, it has become apparent that these faces are nonetheless in significant use in clinical situations, such as in children's hospitals. We have used these base faces to represent up to three intensities for each of the twenty-four emotion categories. Dynamically controlled morphing is used between the approximately sixty faces so that an agent's face may, for example, gradually break into an expression of fear, or immediately scowl with rage. Faces may also decay according to the corresponding rate of decay for the internal emotion state of the represented agent, back to the current default. Defaults for the different agents may represent both dispositional characteristics of the different characters, and current mood. The module runs in real time, allowing for run-time control over face size, rate of morph, color, and rudimentary mouth movement (for when the agent is speaking). The system thus allows for over 3000 different morphs, a range not possible with 3D representation. The morphs run on a 66 Mhz IBM PC (with sound and speech cards) concurrently with midi playback, text-to-speech, speech recognition, and the background emotion simulation, testifying to its computational efficiency.

To keep the graphics module independent we have, in some implementations, also used a simple LL(1) grammar for the control of an agent's facial graphics that allows us to specify the duration of the emotion state, or mood, the intensity, and the rate of decay back to the default state. This design allows us to run the face graphics on a separate machine for each agent, although to date this has not been necessary, and also allows for complete independence in the graphics representation (e.g., it would be trivial to change to a 3D model should a suitable one become available).

4.4 Inflected text-to-speech

We have used a simple, commercially available, text-to-speech module. It was clearly deficient, and we are encouraged by a newer package that we are just now beginning to exploit, although it is too soon to report results on the newer work.[12]

[11] We will appreciate any help in identification.

[12] Since the system runs in real time, we have to adapt all software to our message-passing system, to coordinate it with the facial graphics, to work out conflicts with music generation and speech recognition, etc. We also have to run preliminary tests on the users' ability to identify

Nonetheless, we suggest that even a restricted package has allowed us to support a number of important elements in emotion communication.

First, since the speech stream is dynamic, we have full LISP program control over the construction of utterances, in real time. This means that new words that become available to the lexicon (as discussed above) and expressions based on interactions with the user (such as those derived from past cases retrieved according to remindings provided by the user), can be constructed in the course of dialog with the user, and spoken without having to first be compiled. The automated agent can construct affect-bearing content, with attendant simple explanations, as idiosyncratic states arise within the simulation. For example, an agent might construct the utterance, "I am *very sad* right now Clark, because my-birthday was one month ago, and no-one remembered-my-birthday," based on simple calendar calculations.

Second, dynamic control over speed, pitch, and volume, in even this simple package, allow us to create approximately five distinct inflections that plausibly represent the twenty-four emotion categories. For example, *anger* is represented plausibly as medium fast, loud, and low, while *fear* is slow, high, and medium loud. We also have some small flexibility for expressing emotional intensity.

Third, although this can be rather time-consuming, for static (canned) inflected speech we are able to increase expressivity by inserting inflection control between phonemes, in any phrase we wish.

Fourth, since speech is under LISP program control, if one's standards are not too high, it is trivial to turn text, in for example certain types of stories, into inflected dialog between characters. One such application allows us to create simple interactive stories for children in little more than the time it takes to type in the text. Because it is so easy to add simple affective expression in this way to text-based scenarios, constraints on its applicability to a broad range of appropriate applications are greatly reduced. Enhanced versions, such as those with hypertext explanations, are also relatively economical to create.[13]

Fifth, we have found the speech interface (both text-to-speech, and speech recognition) to be very successful in giving young children, who may not be able to read, access to the computer. Such children are primarily interested in stories, and emotion inflection appears to enhance both understanding, and interest, in the narrations.

---

inflections, to set an upper limit on the number of inflection categories that can be discriminated, to map the emotion categories and intensities to the available inflections, to balance emotion expression with character identification, and so forth. Although we are confident of a significant increase in expressivity, it will take some time to develop the theoretical structures used to take advantage of it.

[13]But note that this simplicity of this application is dependent upon our use of the associated, comprehensive, emotion theory.

5 Closing remarks

In this short article we have attempted to make the case that current technology supports enough rudimentary emotion reasoning, and two-way affective communication with the user, for many years' worth of fruitful study. Since the Affective Reasoner itself has been adapted to run on an inexpensive, widely available, computing platform, it should be apparent that specialized hardware is not required to pursue these avenues of research. We have suggested that humans have untapped communication and social skills that can be utilized to allow complexity to be added to computer applications without endangering the user's comprehension of basic domain tasks.

References

[Bates *et al.*, 1992] Joseph Bates, A. Bryan Loyall, and W. Scott Reilly. Integrating reactivity, goals, and emotion in a broad agent. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, Bloomington, IN, August 1992. Cognitive Science Society.

[Birnbaum and Collins, 1984] Lawrence Birnbaum and Gregg Collins. Opportunistic planning and freudian slips. In *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, Boulder, CO, 1984. Cognitive Science Society.

[Blumberg, 1994] Bruce Blumberg. Building believable animals. In *AAAI Technical Report for the Spring Symposium on Believable Agents*, pages 16–20, Stanford University, March 1994. AAAI, American Association for Artificial Intelligence.

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

[Elliott and Siegle, 1993] Clark Elliott and Greg Siegle. 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*, pages 58–67. American Association for Artificial Intelligence, 1993. Stanford University, March 23-25, Palo Alto, CA.

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

[Elliott, 1993] Clark Elliott. Using the affective reasoner to support social simulations. In *Proceedings of the Thirteenth International Joint Conference on*

Artificial Intelligence, pages 194–200, Chambery, France, August 1993. Morgan Kaufmann.

[Elliott, 1994a] Clark Elliott. Components of two-way emotion communciation between humans and computers using a broad, rudimentary, model of affect and personality. *COGNITIVE STUDIES: Bulletin of the Japanese Cognitive Science Society*, 1(2):16–30, 1994. THIS PAPER. In Japanese with English abstract.

[Elliott, 1994b] Clark Elliott. Multi-media communication with emotion-driven 'believable agents'. In *AAAI Technical Report for the Spring Symposium on Believable Agents*, pages 16–20, Stanford University, March 1994. AAAI, American Association for Artificial Intelligence.

[Elliott, 1994c] Clark Elliott. 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*, pages 9–15, Seattle, WA, August 1994. AAAI, American Association for Artificial Intelligence.

[Elliott, 1994d] Clark Elliott. 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, 1994.

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

[Maes *et al.*, 1994a] Pattie Maes, Trevor Darrel, Bruce Blumberg, Sandy Pentland, and Leonard Foner. Interacting with animated autonomous agents. In *Notes for the Twelfth National Conference on Artificial Intelligence (AAAI-94) Workshop on Artificial Intelligence, Artificial Life, and Entertainment*, pages 43–47. American Association for Artificial Intelligence, August 1994.

[Maes *et al.*, 1994b] Pattie Maes, Trevor Darrel, Bruce Blumberg, and Sandy Pentland. Alive: Artificial life interactive video environment. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, page 1506, Seattle, WA, August 1994. AAAI, American Association for Artificial Intelligence.

[Marquis and Elliott, 1994] Stuart Marquis and Clark Elliott. 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*, pages 11–15. American Association for Artificial Intelligence, August 1994.

[Nagao and Takeuchi, 1994] Katashi Nagao and Akikazu Takeuchi. Social interaction: Multimodal conversation with social agents. In *Proceedings of the Twelfth*

National Conference on Artificial Intelligence, pages 9–15, Seattle, WA, August 1994. AAAI, American Association for Artificial Intelligence.

[Nass and Moon, 1994] Clifford Nass and Youngme Moon. Human-computer interaction as interpersonal interaction: Can computers have personalities? Stanford University. Working title, paper in preparation from interpreted data, 1994.

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

[O'Rorke and Ortony, 1994] Paul O'Rorke and Andrew Ortony. Explaining emotions. *Cognitive Science*, 18:283–323, 1994.

[Ortony *et al.*, 1987] Andrew Ortony, Gerald Clore, and Mark Foss. The referential structure of the affective lexicon. *Cognitive Science*, 11:341–364, 1987.

[Ortony *et al.*, 1988] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[Pelachaud *et al.*, 1993] Catherine Pelachaud, Marie-Luce Viaud, and Hussein Yahia. Rule-structured facial animation system. In *Proceedings of the Thirteenth Annual Joint Conference on Artificial Intelligence*, Chambery, France, August 1993. Morgan Kaufmann.

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

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

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

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

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