Modeling Emotional State and Personality for Conversational Agents

Jack Breese Gene Ball
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
breese@microsoft.com,geneb@microsoft.com

Abstract

We describe an architecture for constructing a character-based agent based on speech and graphical interactions. The architecture uses models of emotions and personality encoded as Bayesian networks to 1) diagnose the emotions and personality of the user, and 2) generate appropriate behavior by an automated agent in response to the user's input. Classes of interaction that are interpreted and/or generated include such things as

- Word choice and syntactic framing of utterances,
- Speech pace, rhythm, and pitch contour, and
- Gesture, expression, and body language.

In particular, we describe the structure of the Bayesian networks that form the basis for the interpretation and generation. We discuss the effects of alternative formulations on assessment and inference.

1 Introduction

Within the human-computer interaction community there a growing consensus that traditional WIMP (windows, icons, mouse, and pointer) interfaces need to become more flexible, adaptive, and human-oriented computer [Flanagan et al., 1997]. Simultaneously, technologies such as speech recognition, text-to-speech, video input, and advances in computer graphics are providing increasingly rich tools to construct such user interfaces. These trends are driving growing interest in agent- or character-based user interfaces exhibiting quasi-human appearance and behavior.

One aspect of developing such a capability is the ability of the system to recognize the emotional state and personality of the user and respond appropriately [Picard, 1995, Reeves and Nass, 1995]. Research has shown that users respond emotionally to their computers. Emotion and personality are of interest to us primarily because of the ways in which they influence behavior, and precisely because those behaviors are communicative– in human dialogues they establish a channel of social interaction that is crucial to the smoothness and effectiveness of the conversation. In order to be an effective communicant, a computer character needs to respond appropriately to these signals from the user and should produce its own emotional signals that reinforce, rather than confuse, its intended communication.

In this paper we address two crucial issues on the path to what Picard has termed *affective computing* [Picard, 1997]:

- Providing a mechanism to infer the likely emotional state and personality of the user, and
- Providing a mechanism to generate behavior in an agent (.e.g. speech and gesture) consistent with a desired personality and emotional state.

2 A Command and Control Agent

We are motivated by a character-based interface to a troubleshooter. Given a malfunctioning device or system, a troubleshooter [Heckerman et al., 1995] generates suggestions for repairs (e.g. reinstall your printing software) or asks for additional information (e.g. what kind of printer do you have?) based on its current state of information. This sequence of actions is driven by a benefit-cost analysis at each stage of the process. Due to the nature of inferential machinery in the system (and unlike a standard decision tree), the user can skip any suggestion or add additional information at any time, and the system will generate a reasonable recommendation. This problem-solving model forms a relatively rich task-oriented dialog between a user (presumably with a faulty device or piece of software) and the system or agent, where the overall direction of the dialog is driven by a myopic cycle of decision-theoretic planning [Heckerman et al., 1995].

Imagine a diagnostic session where a user is having trouble printing and an automated, speech-enabled agent is providing assistance. The agent asks a few informational questions and then makes a suggestion "Please try the following. Go to your printer and make sure all cables are plugged in properly and the printer is turned on and is online." The user checks this and returns, replying "No dice, it still doesn't print." Due to the failure of the speech recognition system to recognize "dice", the agent responds "I'm sorry, I did not understand. Please repeat yourself." The user responds, in a some what faster and louder tone, "I said it didn't work! What should I try next?" The agent, noting the speed, volume, intonation, and wording of the utterance now has an increased probability that the user is upset, and a slightly increased belief that the person is a dominant personality. In response, the agent could decide to be either extremely submissive and apologetic for its failings so far, or respond in kind in a terse, confident fashion. The agent chooses the second path. "Try switching the printer off and back on, and try printing again." it replies, in a somewhat less courteous manner than the previous suggestion.

This dialog is an example of a *command and control* interface, in that at each stage there are relatively few alternatives that the agent (or speech recognizer) needs to consider. In the scenario we are considering, at any point the agent need only consider responses to the previous question, as well as a few generic responses (e.g. quit). As we will see, the recognition and generation of alternative phrasings for these speech acts will provide the basis for an affective infrastructure for the agent.

The architecture we present here is appropriate for a broad range of tasks that are amenable to such command and control interfaces. The architecture does not attempt to manipulate the probabilistic characteristics of the language model used by the speech recognition engine, as proposed in [Horvitz and Shwe, 1995], but rather interprets the various possible rephrasings of a fixed set of alternatives in terms of emotion and personality. In future work, we will be considering the integration of more general natural language grammars into the framework, for both task specification and emotion/personality sensing.

3 Modeling Emotions and Personality

The understanding of emotion and personality is the focus of an extensive psychology literature. In this work, we adopt a simple model in which current emotional state and long term personality style are characterized by discrete values along a small number of dimensions. These internal states are then treated as unobservable variables in a Bayesian network model. We construct model dependencies based on purported causal relations from these unobserved variables to observable quantities (expressions of emotion and personality) such as word choice, facial expression, speech speed, etc.

Bayesian networks are an appropriate tool due to the uncertainty inherent in this domain. In addition, as discussed below, Bayesian network algorithms can perform causal inference (from causes to effects) as well as diagnostic reasoning (from effects to causes), which is directly applicable in this domain. Finally, the flexibility of dependency structures expressible within the Bayes net framework make it possible to integrate various aspects of emotion and personality in a single model that is easily extended and modified.

Emotion is the term used in psychology to describe short-term variations in internal mental state, including both physical responses like fear, and cognitive responses like jealousy. We focus on two basic dimensions of emotional response [Lang, 1995] that can usefully characterize nearly any experience:

- **Valence** represents overall happiness encoded as *positive* (happy), neutral, or *negative* (sad).
- **Arousal** represents the intensity level emotion, encoded as *excited* , *neutral*, or *calm*.

Personality characterizes the long-term patterns of thought, emotion, and behavior associated with an individual. Psychologists have characterized five basic dimensions of personality, which form the basis of commonly used personality tests. We have chosen to model the two traits [McCrae and Costa, 1989] that appear to be most critical to interpersonal relationships:

- **Dominance** indicates a disposition toward controlling or being controlled by others, encoded as *dominant*, *neutral*, or *submissive*.
- **Friendliness** measures the tendency to be warm and sympathetic, and is encoded as *friendly* , *neutral*, or *unfriendly*.

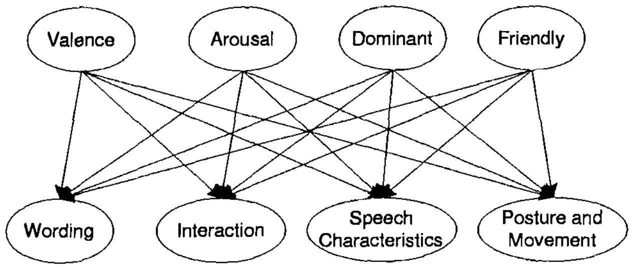


Figure 1: A Bayesian network indicating the components of emotion and personality and various types of observable effects.

Psychologists have devised laboratory tests which can reliably measure both emotional state (with physiological sensing such as galvanic skin response and heart rate) and personality (with tests such as the Myers-Briggs Type Indicator [Myers and McCaulley, 1985]). A computer-based agent does not have these "sensors" at its disposal, so alternative sources of information must be used.

Our Bayesian network therefore integrates information from a variety of observable linguistic and non-linguistic behaviors as shown in Figure 1. Various classes of these observable effects of personality and emotion are shown in the figure. In the following sections, we discuss the range of non-linguistic signals that can be accommodated by our model and then focus in more detail on the way in which the Bayesian network represents the effects of personality and emotion on linguistic expression.

3.1 Non-Linguistic Expression

Humans communicate their emotional state constantly through a variety of non-verbal behaviors, ranging from explicit (and sometimes conscious) signals like smiles and frowns, to subtle (and unconscious) variations in speech rhythm or body posture. Moreover, people are correspondingly sensitive to the signals produced by others, and can frequently assess the emotional states of one another accurately even though they may be unaware of the observations that prompted their conclusions.

The range of non-linguistic behaviors that transmit information about personality and emotion is quite large. We have only begun to consider them carefully, and list here just a few of the more obvious examples. Emotional arousal affects a number of (relatively) easily observed behaviors, including speech speed and amplitude, the size and speed of gestures, and some aspects of facial expression and posture. Emotional valence is signalled most clearly by facial expression, but can also be communicated by means of the pitch contour and rhythm of speech. Dominant personalities might be expected to generate characteristic rhythms and amplitude of speech, as well as assertive postures and gestures. Friendliness will typically be demonstrated through facial expressions, speech prosody, gestures and posture.

The observation and classification of emotionally communicative behaviors raises many challenges, ranging from simple calibration issues (e.g. speech amplitude) to gaps in psychological understanding (e.g. the relationship between body posture and personality type). However, in many cases the existence of a causal connection is uncontroversial, and given an appropriate sensor (e.g. a gesture size estimator from camera input), the addition of a new source of information to our model will be fairly straightforward.

Within the framework of the Bayesian network of Figure 1, it is a simple matter to introduce a new source of information into the model. For example, suppose we incorporate a new speech recognition engine that reports the pitch range of the fundamental frequencies in each utterance (normalized for a given speaker). We could add a new network node that represents PitchRange with a few discrete values, and then construct causal links from any emotion or personality nodes that we expect to affect this aspect of expression. In this case, a single link from Arousal to PitchRange would capture the significant dependency. Then the model designer would estimate the distribution of pitch ranges for each level of emotional arousal, to capture the expectation that increased arousal leads to generally raised pitch. The augmented model would then be used both to recognize that increased pitch may indicate emotional arousal in the user, as well as adding to the expressiveness of a computer character by enabling it to communicate heightened arousal by adjusting the base pitch of its synthesized speech.

3.2 Selection of Words and Phrases

A key method of communicating emotional state is by choosing among semantically equivalent, but emotionally diverse paraphrases– for example, the difference between responding to a request with "sure thing", "yes", or "if you insist". Similarly, an individual's personality type will frequently influence their choice of phrasing, e.g.: "you should definitely" versus "perhaps you might like to".

We have modeled wording choice more deeply than other aspects of the emotion and personality. Since we are concerned with command and control, we have focused on spoken commands. The emotional and personality content is reflected in how someone will ex-

Table 1: Paraphrases for alternative concepts.

Concept	Paraphrases	
greeting	hello hi there howdy	greetings hey
yes	yes yeah I think so	absolutely I guess so for sure
suggest	I suggest that you perhaps you would like to maybe you could	you should let's

press a given concept. Associated with each concept is a set of alternative expressions or paraphrases. Some examples are shown in Table 1.

We model the influence of emotion and personality on wording choice in two stages, as shown in Figure 2. The first stage captures the relationship between personality and emotion and various classes of expressive style. The current model has nodes representing *active*, *positive*, *terse* , and *strong* expression. These nodes are successors of the emotion and personality nodes, and capture the probability that the individual would express themselves in an active, positive, strong, and/or terse manner. Each of these nodes are binary valued, true or false. Thus, this stage captures the degree to which an individual with a given personality and in a particular emotional state will tend to communicate in a particular style.

The second stage captures the degree that each paraphrase actually is active, positive, terse, etc. This stage says nothing about the individual, but rather reflects a general cultural interpretation of each paraphrase, that is the degree to which that phrase will be interpreted as active, positive, terse, and so on by a speaker of American English. A node such as "Concept Active" is also binary valued, and is true if the paraphrase would be interpreted as "Active" and false otherwise.

Finally there is a set of nodes representing whether a particular expressive interpretation of the paraphrase matches the intended expressive style by the individual for each component. A node such as "Match Active" has value true if and only if the values of "Concept Active" and "Active Expression" are the same. The node "Match" at the bottom of the network is simply a Boolean conjunction that has value *true* when all its parents (the match nodes for each component of expressive style) are true.

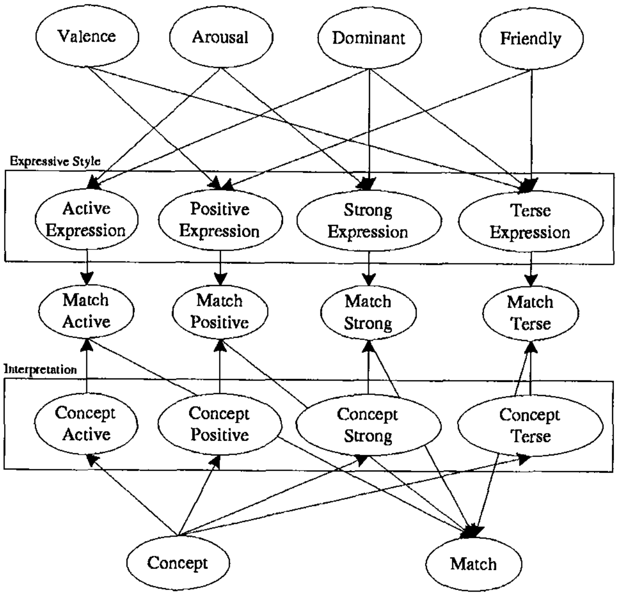


Figure 2: A belief network fragment indicating 1) the relationship of emotion and personality on expressive style 2) the probability that a modeled concept will be interpreted as a particular style, and 3) whether the interpretation matches the intent for each component and whether they match on all components.

In developing a version of this Bayes net for a particular application (such as troubleshooting), we need to generate a network fragment such as shown in Figure 2 for each possible conceptual command or element in the vocabulary of the application. These fragments are merged into a global Bayesian network capturing the dependencies between the emotional state, personality, natural language, and other behavioral components of the model. A portion of such a merged Bayesian network is shown in Figure 3.

The various fragments differ only in the assessment of the paraphrase scorings, that is the probability that each paraphrase will be interpreted as active, strong, etc. There are five assessments needed for each alternative paraphrase for a concept (the ones mentioned earlier, plus a formality assessment). Note that the size of the belief network representation grows linearly in the number of paraphrases (the number of concepts modeled times the number of paraphrases per concept).

In a previously proposed model structure, we had each of the expressive style nodes pointing directly into the concept node, creating a multi-stated node with five parents. The assessment burden in this structure was substantial, and a causal independence assumption such as noisy-or is not appropriate [Heckerman, 1993]. The current structure reduces this assessment burden,

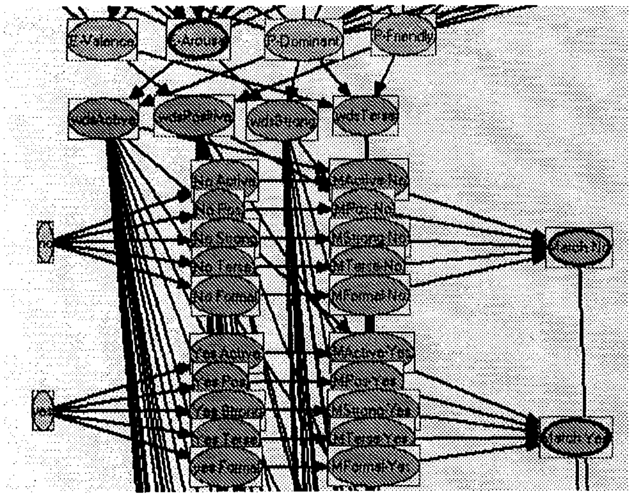
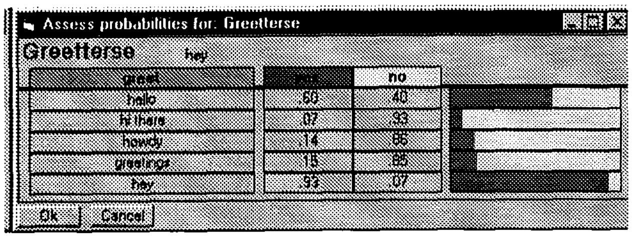


Figure 3: A portion of a merged Bayesian network for emotion/personality/vocabulary.


Assess probabilities for: Greetterse

Greetterse hey

greet	yes	no
hello	.60	.40
hi there	.07	.93
howdy	.14	.86
greetings	.15	.85
hey	.93	.07

Figure 4: An example assessment window for the probabilities that the various paraphrasing of "greet" will be perceived as "terse".

and also allows modular addition of new "expressive style" nodes. If we add a new expressive style node to the network (such as "cynical"), then the only additional assessments we need to generate are the "cynical" interpretation nodes of each concept paraphrase. An assessment of this type is shown in Figure 4. These features of the Bayes network structure make it easy to extend the model for new concepts and dimensions of expressive style.

4 Inference

As discussed above, the merged Bayesian network model relates emotional state and personality to wording choice, speech characteristics, input style characteristics, and body language/movements. Most of these observable expressions are modeled as being directly caused by the the components of emotion and personality. For choice of paraphrase we make an additional assumption in using the Bayes net structure described above: the individual being modeled choose wording so as to match the intended interpretation with their current desired expressive style. Thus we are imputing some choice behavior to the individual. This behavior is incorporated into inference by setting the "match" nodes to true before updating probabilities. Students of Bayesian inference will note that in the network in Figure 2 observing "match" will serve to create the desired dependency between the components of emotion and personality and the choice of paraphrase.

Under this interpretation, the model captures a decision model regarding word selection. The selection of a paraphrase is done such that it maximizes the probability of a match between intended expressive style and interpretation, given all previous observations regarding gesture, speech characteristics, and wording choice. We implement this approach in the network by setting each "Match" node to true, using the idea originally proposed by Peot and Shachter [Shachter and Peot, 1992]. By setting the prior probability of the paraphrases in each concept node to a uniform distribution over the alternatives, application of a standard Bayesian network inference algorithm will generate a posterior distribution over word choices consistent with "match" being true. The paraphrase that has the maximum posterior probability is the one that maximizes the probability of "match" being true. We discuss the use of this technique more fully in the next section where we describe using this model for diagnostic (What mood is the user in?) as well as for generating behavior (What should the agent say if he is in a good mood?).

There are additional issues relating to the dynamic construction of this linguistic paraphrase model that have not been fully explored. Currently we construct the network fragments for each concept (as shown in Figure 2) and include all possible concepts in a static merged network as in Figure 3. It may be more efficient to construct the model dynamically, while performing inference, for those items that one wishes to diagnose or generate. This approach was explored in a preliminary fashion in [Goldman and Breese, 1992].

5 Reasoning Architecture

In the agent we maintain 2 copies of the emotion/personality model. One is used to diagnose the user, the other to generate behavior for the agent. The basic setup is shown in Figure 5. In the following, we will discuss the basic procedures used in this architecture, referring to the numbered steps in the figure.

1. Observe. This step refers to recognizing an utterance as one of the possible paraphrases for a

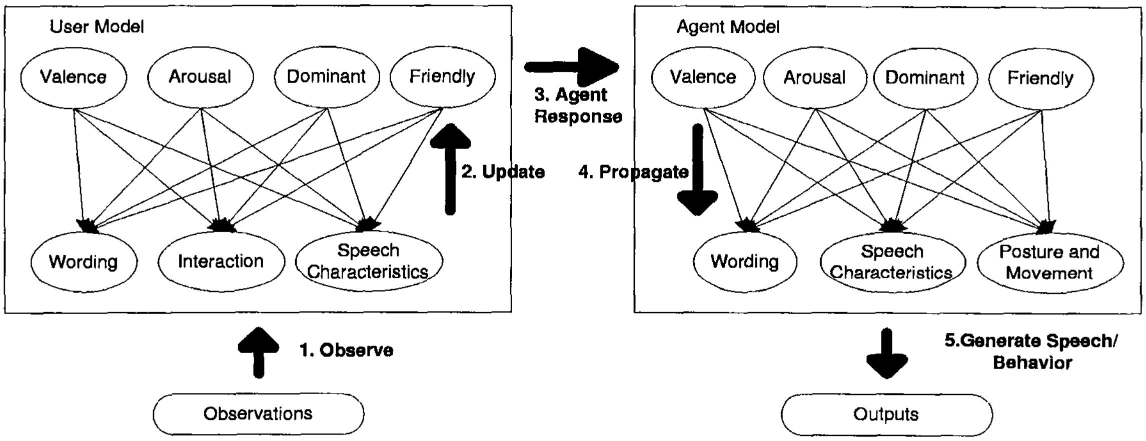


Figure 5: An architecture for speech and interaction interpretation and subsequent behavior generation by a character based agent.

concept. At a given point in the dialog, for example after asking a yes/no question, the speech recognition engine is listening for all possible paraphrases for the speech concepts *yes* and *no*. When one is recognized, the corresponding node in the user Bayesian network is set to the appropriate value.

2. Update. Here we use a standard probabilistic inference [Jensen, 1996, Jensen et al., 1989] algorithm to update probabilities of personality and emotional state given the observations. As discussed in Section 4, all concept "match" nodes are set to true.

3. Agent Response. The linkage between the models is captured in the agent response component. This is the mapping from the updated probabilities of the emotional states and personality of the user to the emotional state and personality of the agent. The response component can be designed to develop an *empathetic* agent, whose mood and personality matches that of the user, or a *contrary* agent, whose emotions and personality tend to be the exact opposite of the user. Research has indicated that users prefer a computerized agent to have a personality makeup similar to their own [Reeves and Nass, 1995], so by default our prototypes implement the *empathetic* response policy. This policy was also chosen by the agent in our example interaction in Section 2.

4. Propagate. Again we use a probabilistic inference algorithm to generate probability distributions over paraphrases, animations, speech characteristics, and so on, consistent with the emotional state and personality set by the response module.

5. Generate Behavior- At a given stage of the dialog, the task model may dictate that the agent express a particular concept, for example *greet* or *regret*. We then consult the agent Bayesian network for the current distribution over the possible paraphrases for expressing that concept. We can chose that paraphrase with the maximum probability, or sample from that distribution to select a particular paraphrase. This string is then passed to the text-to speech engine for generating output. Similar techniques are used to generate animations, and adjust speech speed and volume, or other affect-sensitive actions.

6 Conclusions and Future Work

This paper presents work in progress in developing adaptive conversational user interfaces. We have presented a reasoning architecture for an agent that can recognize a user's personality and emotional state and respond appropriately in a non-deterministic manner. The model and architecture have been implemented in two prototypes at Microsoft Research. In future studies, we plan to validate the predictions of the models by comparing them to human judgment in diagnosing emotion and personality.

As far as application of the model in computer systems, the current model applies to a command and control scenario where there are relatively few utterances and responses that the agent needs to recognize and respond to. In future work we would like to address the more general case of dictation-style speech recognition with more complete language grammars.

The current model has no explicit notion of time, obviously a deficiency when modeling emotions and interaction. The existing model can be extended to a dynamic Bayesian network [Dean and Kanazawa, 1988]. The temporal structure of emotional states would be captured in a temporal lag structure in the dynamic Bayesian network, equivalent to the hidden Markov models posited by Picard [Picard, 1995, Picard, 1997]. The word choice structure in each time period would vary based on the history of words recognized. The system would be able to predict current and future emotional states based on a history of interaction [Shafer and Weyrath, 1997].

References

[Dean and Kanazawa, 1988] Dean, T. and Kanazawa, K. (May, 1988). Probabilistic temporal reasoning. Technical report, Brown University.

[Flanagan et al., 1997] Flanagan, J., Huang, T., Jones, P., and Kasif, S., editors (1997). *Final Report of the NSF Workshop on Human-Centered Sysetem:Information, Interactivity, and Intelligence*, Washington, D.C. National Science Foundation.

[Goldman and Breese, 1992] Goldman, R. P. and Breese, J. (1992). Integrating model construction and evaluation. In *Proceedings of Eighth Conference on Uncertainty in Artificial Intelligence*, Stanford, California. Morgan Kaufmann.

[Heckerman, 1993] Heckerman, D. (1993). Causal independence for knowledge acquisition and inference. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, pages 122–127. Morgan Kaufmann.

[Heckerman et al., 1995] Heckerman, D., Breese, J., and Rommelse, K. (1995). Troubleshooting under uncertainty. *Communications of the ACM.*

[Horvitz and Shwe, 1995] Horvitz, E. and Shwe, M. (1995). Melding Bayesian inference, speech recognition, and user models for effective handsfree decision support. In *Proceedings of the Symposium on Computer Applications in Medical Care.* IEEE Computer Society Press.

[Jensen, 1996] Jensen, F. V. (1996). *An Introduction to Bayesian Networks.* Springer-Verlag, New York, New York.

[Jensen et al., 1989] Jensen, F. V., L., L. S., and G., O. K. (1989). Bayesian updating in recursive graphical models by local computations. Technical Report Report R 89-15, Institute for Electronic Systems, Department of Mathematics and Computer Science, University of Aalborg, Denmark.

[Lang, 1995] Lang, P. (1995). The emotion probe Studies of motivation and attention. *American Psychologist,* 50(5):372–385.

[McCrae and Costa, 1989] McCrae, R. and Costa, P.T., J. (1989). The structure of interpersonal traits: Wiggin's circumplex and the five-factor model. *Journal of Personality and Social Psychology,* 56(5):586–595.

[Myers and McCaulley, 1985] Myers, I. B. and McCaulley, M. H. (1985). *Manual: A Guide to the development and use of the Myers-Briggs Type Indicator.* Consulting Psychologists Press, Palo Alto, CA.

[Picard, 1995] Picard, R. W. (1995). Affective computing. Technical Report 321, M.I.T. Media Lab Perceptual Computing Section, Cambridge, MA.

[Picard, 1997] Picard, R. W. (1997). *Affective Computing.* MIT Press, Cambridge, MA.

[Reeves and Nass, 1995] Reeves, B. and Nass, C. (1995). *The Media Equation.* CSLI Publications and Cambridge University Press, New York, New York.

[Shachter and Peot, 1992] Shachter, R. and Peot, M. (1992). Decision making using probabilistic inference methods. In *Proceedings of Eighth Conference on Uncertainty in Artificial Intelligence*, Stanford California. Morgan Kaufmann.

[Shafer and Weyrath, 1997] Shafer, R. and Weyrath, T. (1997). Assessing temporally variable user properties with dynamic Bayesian networks. In Jameson, A., Paris, C., and Tasso, C., editors, *User Modeling: Proceedings of the Sixth International Conference*, New York. Springer-Verlag Wien.