What Would They Think?
A Computational Model of Attitudes

Hugo Liu
MIT Media Laboratory
20 Ames St., Cambridge, MA
hugo@media.mit.edu

Pattie Maes
MIT Media Laboratory
20 Ames St., Cambridge, MA
pattie@media.mit.edu

ABSTRACT
A key to improving at any task is frequent feedback from people whose opinions we care about: our family, friends, mentors, and the experts. However, such input is not usually available from the right people at the time it is needed most, and attaining a deep understanding of someone else’s perspective requires immense effort. This paper introduces a technological solution.

We present a novel method for automatically modeling a person’s attitudes and opinions, and an opportunistic interface agent called “What Would They Think?” which offers the just-in-time perspectives of people whose opinions we care about, based on whatever the user happens to be reading or writing. In the application, each person is represented by a “digital persona,” generated from the automated analysis of personal texts (e.g. weblogs and papers written by the person being modeled) using natural language processing and commonsense-based textual-affect sensing.

In user studies, participants using our application were able to grasp the personalities and opinions of a panel of strangers more quickly and deeply than with either of two baseline methods. We discuss the theoretical and pragmatic implications of this research to intelligent user interfaces.

Categories and Subject Descriptors
H.5.2 [User Interfaces]: interaction styles, natural language;
I.2.7 [Natural Language Processing]: language models, text analysis.

General Terms

Keywords
Affective interfaces, affective memory, user modeling.

1. INTRODUCTION
Have you ever been engaged in a task – whether it’s reading the news, writing a research paper, or pursuing a new hobby – where you felt uncertain about how to interpret a situation, and you thought, “what would they think?” This is a common experience, because observing and modeling the attitudes and emotional reactions of others is an important aspect of how humans learn (Bandura, 1977).

Perhaps this is why frequent and timely feedback from people whose opinions we value – our family, friends, mentors, or experts – aids our ability to interpret situations and make decisions. However, we often lack access to the people whose feedback we value, so we are forced to learn about their perspectives in other ways, e.g. by inferring attitudes and opinions from their books and papers. Forming a deep understanding of a person in this manner requires immense effort, and there is no guarantee that we can recall a person’s opinion on a specific topic at the time we need that feedback the most.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Figure 1. A panel of virtual AI researchers offer their attitudes on a passage of text that the user is reading. A green-tinted face indicates a pleasurable response, while a red-tint indicates unpleasant. Brightness corresponds to affective arousal over a topic.

This paper introduces a technological solution to the problem of getting the just-in-time perspectives of people whose opinions we care about. We have built a system that can automatically generate a model of a person’s attitudes and opinions from an automated analysis of a corpus of personal texts written by the person being modeled, consisting of, inter alia, weblogs, emails, editorial papers, and transcribed speeches. The interface agent “What Would They Think?” (WWTT) (Fig. 1) displays a handful of these digital personas together, each reacting affectively to whatever the user happens to be reading or writing. Personas are also...
capable of explaining why they react as they do, with salient quotes from their personal text to justify an affective perspective.

To build a digital persona, the attitudes that a person exhibits in his/her personal texts are recorded into an affective memory system. Newly presented text triggers memories from this system and forms the basis for an affective reaction. Mining attitudes from text is achieved through natural language processing, commonsense-based textual affect sensing (Liu et al., 2003), and the affective memory system. This approach to person modeling is quite novel when compared to previous work on the topic (behavior modeling, e.g. (Sison & Shimura, 1998), and demographic profiling, e.g. questionnaire-derived user profiles).

A related paper on this work (Liu, 2003b) presents a cognitive science perspective and gives a more thorough technical treatment of the work. This paper does not dwell on the implementation-level details of the system, but rather, describes the computational model of attitudes in a more practical light, and discusses how these models are incorporated to build the intelligent user interface “What Would They Think?”.

This paper is structured as follows. First, we introduce a computational model of a person’s attitudes, a system for automatically acquiring this model from personal texts, and methods for applying this model to predict a person’s attitudes. Second, we present how a collection of digital personas can portray a community in “What Would They Think?” and an evaluation of our approach. Third, we situate our work in the literature. The paper concludes with further discussion and presents directions for future work.

2. COMPUTING A PERSON’S ATTITUDES

First-person texts such as, inter alia, weblog diaries, emails, editorial papers, and transcribed speeches and interviews, are rich sources of attitudes and opinions. People are very good at inferring attitudes from text and compiling them into a model of the author, but computationally the problem is more challenging.

Our approach to the problem can be summarized as follows. We implement a computer reader to skim a corpus of personal texts and judge the affect of the text at the document, paragraph, sentence, and concept level. For this task, we use a commonsense-based textual affect sensing engine, described in (Liu et al., 2003), because its robustness relies on having a large corpus of affective commonsense knowledge, not on the presence of obvious mood keywords. A single instance of an attitude is computed as an affective valence score associated with a concept, topic, or “episode.” Instances of attitudes are mined from each personal text, and accumulate in an affective memory system. The affective memory has a reflexive component, which uses single instances of attitudes to condition (in the traditional behaviorist sense of the word) the formation of a more stable attitude. This feature helps make our attitudes model more robust, as the affect sensing engine may not classify every instance of a concept correctly. It is known that some linguistic devices will fool the engine, such as sarcasm, and complex clauses. Conditioning allows only stable and consistently classifiable attitudes to emerge.

In this section, we first present a bipartite model of the affective memory system. Second, we describe the mechanism for mining attitudes from personal texts. Third, we discuss how an affective memory system can be applied to predict a person’s affective reaction to new texts. Fourth, we describe how some advanced features enrich our basic person modeling approach.

2.1 A Bipartite Affective Memory System

A person’s affective reaction to a concept, topic, or situation can be thought of as either instinctive, due to attitudes and opinions conditioned over time, or reasoned, due to the effect of a particularly vivid recalled memory. Borrowing from cognitive models of human memory function, attitudes that are conditioned over time can be best seen as a reflexive memory, while attitudes resulting from the recall of a past event can be represented as a long-term episodic memory (LTEM). Memory psychologist Endel Tulving equates LTEM with “remembering” and reflexive memory with “knowing” and describes their functions as complementary (Tulving, 1983). We combine the strengths of these two types of memory to form a bipartite, episode-reflex model of the affective memory system.

2.1.1 Affective long-term episodic memory

Long-term episodic memory (LTEM) is a stable memory capturing significant experiences and events. The basic unit of memory, called an episode, captures a coherent series of sequential events. Episodes are content-addressable, meaning, they can be retrieved through a variety of cues encoded in the episode, such as a person, location, or action. With LTEM, even events that happen only once can become salient memories and can recurrently influence a person’s future thinking. In modeling attitudes, we must account for the influence of these particularly powerful one-time events.

In our affective memory system, we compute an affective LTEM as an episode frame, coupled with an affect valence score that best characterizes that episode. In Fig. 2, we show an episode frame for the following example episode: “John and I were at the park. John was eating an ice cream. I asked him for a taste but he refused. I thought he was selfish for doing that.”

<table>
<thead>
<tr>
<th>:::: EPISODE FRAME ::::</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUBEVENTS:</strong></td>
</tr>
<tr>
<td>(eat John “ice cream”),</td>
</tr>
<tr>
<td>(ask John “for taste”),</td>
</tr>
<tr>
<td>(refuse John)</td>
</tr>
<tr>
<td><strong>MORAL:</strong> (selfish John)</td>
</tr>
<tr>
<td><strong>CONTEXTS:</strong> (date), (park), ()</td>
</tr>
<tr>
<td><strong>EPISODE-IMPORTANCE:</strong> 0.8</td>
</tr>
<tr>
<td><strong>EPISODE-AFFECT:</strong> (-0.8,0.7,0)</td>
</tr>
</tbody>
</table>

Figure 2. An episode frame in affective LTEM.

As illustrated in Fig. 2, an episode frame decomposes the text of an identified and parsed episode into simple verb-subject-argument propositions like (eat John “ice cream”). Together, these constitute the subevents of the episode. The “moral,” or root cause, of an episode is important because the episode-affect can be most directly attributed to it. The details of extracting morals are presented elsewhere (Liu, 2003b).

The affect valence score in the above example is a numeric triple representing valences in the three nearly independent affective dimensions of Pleasure-Displeasure (i.e., feeling happy or unhappy), Arousal-Nonarousal (i.e., arousing one’s feelings), and Dominance-Submissiveness (i.e., the amount of confidence/lack-of-confidence felt). This is known as the PAD model (Mehrabian, 1995) for short. Each dimension can assume values from –100% to +100%, and a PAD valence score is a 3-tuple of these values (e.g. [–51, .59, .25] might represent anger). The robustness implications of PAD’s continuous account of affect makes it preferable to finite repertoire models of affect such as Manfred Clyne’s
2.1.2 Affective reflexive memory

While long-term episodic memory deals in salient, one-time events and must generally be consciously recalled, reflexive memory is full of automatic, instant, almost instinctive associations. Reflexive memories are formed through the conditioning of repeated exposures rather than one-time events. The conditioning process also acts as a natural noise filter against any incorrect textual affect classifications.

The affective reflexive memory is represented by a lookup-table. The lookup-keys are simple concepts which can be semantically recognized as a person, action, object, activity, or topic. Associated with each key is a list of exposures, where each exposure represents a distinct instance of that concept appearing in the personal texts. An exposure, E, is represented by the triple: (date, affect valence score V, saliency S). At runtime, the affect valence score associated with a given conceptual cue can be computed using the formula given in Eq. (1).

\[
\frac{1}{n} \left[ \log_b \left( \max(n, b) \right) \right] \sum_{t=startdate}^{enddate} S(E_t) V(E_t) \]

(1)

where \( n \) = the number of exposures of the concept; \( b = 2 \)

This formula returns the valence of a conceptual cue averaged over a particular time period. The term, \( \left[ \log_b \left( \max(n, b) \right) \right] \), rewards frequency of exposures, while the term, \( S(E_t) \), rewards the saliency of an exposure. In this simple model of an affective reflexive memory, we do not consider phenomena such as belief revision, reflexes conditioned over contexts, or forgetting.

To give an example of how affective reflexive memories are acquired from personal texts, consider Fig. 3, which shows two excerpts of text from a weblog and a snapshot sketch of a portion of the resulting reflexive memory.

**Text Excerpts**

...2 Oct 01... Telemarketers harassed me today, interrupting my dinner. I'm really upset...

...4 Oct 01... The phone rang, and of course, it was a telemarketer. Damn it!

**Figure 3.** How reflexive memories get recorded from excerpts.

In the above example, two text excerpts are processed with textual affect sensing and concepts, both simple (e.g. telemarketer, dinner, phone) and compound (e.g. telemarketer::call, interrupt::dinner, phone::ring) are extracted. The saliency of each exposure is determined by heuristics such as the degree to which a particular concept in topicalized in a paragraph. The resulting reflexive memory can be queried using Eq. (1). Note that while a query on 3 Oct 01 for “telemarketer” returns an affect valence score of (-1.5, 25.1), a query on 5 Oct 01 for the same concept returns a score of (-2.4, 29.11). Recalling that this valence triple corresponds to (pleasure, arousal, dominance), we can interpret the second annoying intrusion of a telemarketer’s call as having conditioned a further displeasure and a further arousal to the word “telemarketer”.

How does conditioning help the system cope with noise? In Fig. 3, “phone” also inadvertently inherits some negative affect. However, unless “phone” consistently appears in a negative affective context, Eq. (1) will tend to cancel out inconsistent affect valence scores, resulting in a more neutral valence.

In summary, we have motivated and characterized the two components of the affective memory system: an episodic component emphasizing the affect of one-time salient memories, and a reflexive component, emphasizing instinctive reactions to conceptual cues that are conditioned over time. In the following subsection, we propose how this bipartite affective memory system can be acquired automatically from personal texts.

### 2.2 Mining Attitudes from Personal Texts

The bipartite affective memory system presented above is the framework we use to represent a person’s attitudes, and the main method for affective appraisal of new textual episodes is based on triggered memories from this attitudes model.

Other affective modeling frameworks in the literature such as (Gratch & Marsella, 2001) offers much more cognitively sophisticated models of affect generation through the interaction between beliefs, desires, and goals. These models can generate emotional behavior and personality that is more complex and dynamic than the deterministic affective reactions presently generated by our simple memory-based affective appraisal model. The complexity of our present model is constrained because robustly inferring beliefs, desires, and goals from unconstrained text is far more difficult than even inferring personal affect from text. Therefore, our model represents only an initial exploration of how personal attitudes can be modeled from unconstrained texts.

The process of mining attitudes from personal texts to populate the affective memory system involves natural language processing to deconstruct text into concepts (for the reflexive memory), and episodes (for the episodic memory), and textual affect sensing to assign affect scores to each concept and episode. In this subsection, we focus mainly on the affective appraisal problem, and discuss fitness criteria for the inputted personal texts.

### 2.2.1 Affective Appraisal of Personal Text

Appraising the affect of personal text is a difficult task. The affect classification method needs to be able to judge affect at the sentence-level with good accuracy. Several common approaches fail to meet the criteria. Naive keyword spotting looks for mood keywords, but this is insufficient as a stand-alone method because affect is often conveyed without mood keywords. Statistical affect classification using statistical learning models such as latent semantic analysis (Deerwester et al., 1990) generally require large inputs and thus, cannot appraise texts with very much granularity.

To analyze personal text with the desired robustness, granularity, and specificity, we employ a model of textual affect sensing using real-world knowledge, proposed by Liu et al. (2003). In this model, defeasible knowledge of everyday people, things, places, events, and situations from the Open Mind Commonsense (OMCS) corpus (Singh et al., 2002) is leveraged to sense the affect of a text by evaluating the affective implications of each event or situation. For example, to evaluate the affect of “I got
fired today,” this model evaluates the consequences of this situation and characterizes it using negative emotions such as fear, sadness, and anger. This model, coupled with a naïve keyword spotting approach, provides rather comprehensive and robust affective classification. The output of the textual affect sensing subsystem is a PAD score.

One of the most interesting issues is learning personal affect using a person-neutral affect sensing mechanism. Since the OMCS corpus was built collaboratively by 11,000 web teachers, the assessment made by such a model represents the judgment of a typical person, which may sometimes be different from the affective judgment of the particular person being modeled. However, we can assume that although a personal affect judgment may deviate from that of a typical person on small particulars, it will not deviate on average, when examining a large text. The implication of this is that on a slightly larger granularity than a sentence, the affective appraisal is more likely to be accurate. In fact, accuracy should increase proportional to the size of the textual context being considered. The evaluation of Liu et al.’s affective navigation system (2003b) yields some indirect support for the idea that accuracy increases with the size of the textual context. In that user study, users found affective categorizations of textual units on the order of chapters to be more accurate and useful to information navigation than affective categorizations of small textual units such as paragraphs.

To assess the affect of a sentence, we factor in the affective assessment of not only the sentence itself, but also of the paragraph, section, and whole journal entry or episode. Because so much context is factored into the affect judgment, only a modest amount of affective information can be learned for any given sentence. Thus we rely on the confirming effects of being able to encounter an attitude multiple times (i.e. conditioning the reflexive memory). In exchange for only being able to learn a modest amount from a sentence, we minimize the impact of erroneous judgments.

### 2.2.2 What Personal Texts are Suitable?

Suitable personal texts satisfy the following criteria. 1) Texts should be first-person, because having to attribute opinions to multiple sources would require more in-depth story understanding. 2) Texts should be rich sources of candid opinion, so an editorial paper would be better than a dispassionate paper. 3) If the digital persona is intended to represent the whole of a person’s personality and attitude, the selection of personal texts should cover a good breadth of topics, not covering one or two topics disproportionately. Unbalanced collections of personal texts will generate digital personas skewed toward particular discourses, and texts that are too esoteric will hurt the accuracy of the commonsense-based appraisal mechanism. 4) Text sources like a weblog diary is preferred because it covers attitudes and opinions on day-to-day life, and has an explicit episodic organization. In texts not already organized by episodes, the natural language processing mechanism will try to heuristically segment the text.

Even if these fitness criteria for personal texts are met, there are inherent limitations to the affective appraisal mechanism. 1) Since independent clauses are the largest units of context addressed by the commonsense-based affect sensing engine, concepts within simple declarative assertions can be appraised much more accurately than concepts within complex arguments. The appraisal of “Mr. X” in the following passage would be erroneously positive: “Mr. X is such a nice guy. Everyone loves Mr. X. Gimme a break!” 2) The affective appraisal mechanism cannot recognize humor and sarcasm because it does not have much knowledge about expectation-violation. It does not know that if you say “I hope you die a horrible death” to a friend who has just played a joke on you, it was probably meant as an exaggeration.

The system has two coping strategies for erroneous affect appraisals. First, as discussed in section 2.2.1, the affective assessment of a sentence is tempered by the affect of the paragraph, section, and document in which it is contained, since larger contexts of texts can be appraised with less noise. Second, the process of conditioning in the formation of reflexive memories will tend to cancel out inconsistent instances of affective appraisal.

In summary, digital personas can be automatically acquired from suitable personal texts using natural language processing and textual affect sensing. Suitable texts meet certain fitness criteria such as being first-person, opinion-rich, well-balanced, and episodic-organized. The proposed affective appraisal mechanism employs coping strategies for dealing with erroneous appraisals, especially over sarcastic, humorous, or argument-based text.

### 2.3 Predicting Attitudes using the Model

Having acquired the model, the digital persona attempts to predict the attitudes of the person being modeled by offering some affective reaction when it is fed a new textual episode. The new textual episode is parsed into objects and associated attributes. For example, “Computers are dumb” will return an object attribute pair, “(computers, dumb)”. The list of attributes is affectively appraised using commonsense-based textual affect sensing and a back-off mood keyword spotter, and an attitude is generated: “(computers, -3)”. The second term -3 is the “P” dimension of resulting PAD score. Objects in the text, as well as the jisted topics of the text are compared with the concepts in the affective memory system. If an object in the text triggers the object in memory, the Arousal and Dominance valence scores are attached to the new textual episode. The Pleasure valence score is compared with the attitudes in the new textual episode. If the attitudes can be aligned, the affective reaction is one of agreement. If the attitudes expressed in the new textual episode oppose the attitudes expressed in the memory, the affective reaction is one of disagreement. The gestalt reaction to the new text sums up the contributory effects of each triggered memory.

The triggering process for episodic memory is somewhat more complex, requiring the detection of an episode in the new text, and heuristically pattern matching this new episode frame to the library of episode frames. The range of concepts that can trigger a reflex memory (i.e. fuzzy triggering) is increased by the addition of conceptual analogy. The details are omitted here, but are discussed elsewhere (Liu, 2003b).

Predicting a person’s affective reaction to new text is likely to be error prone if only one or two triggered memories account for the prediction. This is because each affective memory only contributes a small piece of context to the prediction. Successful prediction relies on the premise that the contexts of many triggered memories will have some commonality and overlap, and this contextual intersection will lead to a better prediction.

### 2.4 Enriching the Basic Model

The basic model of a person’s attitudes uses attitudes mined from personal texts to appraise new textual episodes. While this basic model is sufficient to produce reactions to text for which there
exists some relevant personal memories, the generated digital personas are often quite “sparse” in what they can react to. We have proposed and evaluated some advancements to the basic model. In particular, we have looked at how a person’s attitude model can be enriched by the attitude models of people whom the modeled person fashions himself/herself after – perhaps a good friend or mentor. More technically, we mean an imprimer. Marvin Minsky describes an imprimer as someone whose goals and attitudes we admire and hope to emulate. From the supposition that we aspire to many of the attitudes of our imprimers, we hypothesize that affective memory models of these imprimers, if known, can complement the person’s own affective memory model in helping to predict a person’s attitudes. This hypothesis is supported by work in psychoanalysis (Freud, 1991) on attitude introjection. Based on Minsky’s suggestion that imprimers evoke self-conscious emotions like pride and embarrassment, we developed and implemented a heuristic approach to automatically identifying imprimers from a person’s affective memory. Once identified, the system searches for text on the imprimer and attaches the imprimer’s affective memory model to supplement the person’s own affective memory when appraising new textual episodes. (Liu, 2003b) provides a more complete account of imprimers in attitude modeling. Indicative trials suggest that imprimers can enhance the performance of affective appraisal of new texts.

In summary, we have presented a reflex-episode model of affective memory as a memory-based representation of a person’s attitudes. The model can be acquired automatically from personal text using natural language processing and textual affect analysis. The model can be applied over new textual episodes to produce affective reactions that aim to emulate the actual reactions of the person being modeled. We have also discussed how the basic attitudes model can be enriched with added information about the attitudes of the mentors of the person being modeled.

In the following section, we describe how digital personas are composed to create the What Would They Think? application.

3. WHAT WOULD THEY THINK?

What Would They Think? (WWTT) is an opportunistic interface agent which offers the just-in-time perspectives of people whose opinions we care about, based on whatever the user happens to be reading or writing. (Fig. 1). The application consists of a panel of “advisors” who sit on the desktop. An interface agent observes a user as he/she browses a webpage, writes an essay, or replies to an email, and the advisors constantly react to the current text being read or written. In this section, we present elements of the interface design, followed by discussions of two evaluations, one for the underlying attitudes model and one for the application.

3.1 Interface Design

Digital personas acquired from an automatic analysis of personal text, are represented visually with pictures of faces, which occupy a panel (or n x n matrix, to accommodate more personas). Given a new textual episode, each persona expresses an affective reaction by modulating the graphical elements of its icon. Each digital persona is also capable of explaining what motivated its reaction by displaying salient quotes from its repository of personal texts.

The Iconographic Face. A virtual representation of a person is given as a normalized, gray-scaled image of that person’s face. Affective reactions are conveyed through modulations in the color, brightness, and sharpness of the face image. From early experimentation, we found that faces are a far more convincing visual metaphor for a digital persona than something textual or abstract. People are pre-wired with the ability to quickly recognize and remember faces, and to use a face as a cognitive container for an individual’s unique identity and personality.

We deliberately chose to convey affect through modulations of the image rather than through realistic manipulations of facial expression and gaze. It is important to not portray more detail in the face than our attitude model is capable of elucidating, for the face is fraught with social cues, and unjustified cues could do more harm than good. Scott McCloud has explored extensively the representational-vs.-realistic tradeoff of face drawing in comics (1993).

Visualizing an affective reaction. We employ a rather straightforward scheme to map the three PAD dimensions of an affective reaction (pleasure, arousal, dominance) onto the three graphical dimensions of color, brightness and sharpness, respectively. The baseline image being modulated is gray-scaled and its brightness and contrast are equalized to be uniform across all images. Using a traffic light metaphor, a pleasurable or approving reaction tints a face green, while an unpleasurable / disapproving reaction tints a face red. An affectively aroused reaction results in a brightly lit icon, while a non-aroused reaction results in a dimly lit icon. A dominant (confident) reaction maps to a sharp, crisp image, while a submissive (unconfident) reaction maps to a blurry image. While better mapping schemes may exist, our experience with users who have worked with this interface tells us that the current scheme conveys the affect reaction quite intuitively.

Configuring the Panel. Presently, WWTT is configured for several tasks. The user can use WWTT to get the just-in-time perspectives of a panel of advisors reacting to text read and typed. To add a persona to the panel, a user specifies a face icon, and a url to a weblog or to a corpus of texts, which obey the suitability criteria given in section 2.2.2 (although this is not explicitly enforced). In addition, WWTT can be used to visualize the personalities and strong opinions of an online community. The application automatically analyzes any one of several online communities – including a blog ring, a circle of friends on friendster.com, and a usenet newsgroup – and generates an appropriate matrix of personas. WWTT has also been implemented to react to the content of conversations using speech recognition.

Explanation. A digital persona is capable of some limited explanation. Clicking on a persona’s reaction will display a collection of salient quotes from that persona’s text. These quotes are generated by backpointers to the text associated with each affective memory. For episodic memory, a particularly salient episode is quoted, while there are many quotes given to support a triggered reflex memory.

The presentation of the quotes is ordered by saliency and relevance. Quotes which make the largest contribution to the resulting affective reaction are promoted to the top of the explanation page. Similarly, quotes which most exemplify the affective reaction are promoted to the top.

In most cases, triggered quotes can only offer indirect and partial justification for a persona’s reaction to a new textual episode because the new episode creates a new interpretational context governing the meaning of each concept triggered in memory. However, with exercising some critical thinking and synthetic
reasoning, a user should be able to verify from the indirect explanation whether or not an affective reaction is indeed justified. This lends the interface some fail-softness, as a user can recover if the system erroneously represents a person’s reaction.

3.2 Evaluation of Underlying Attitude Models
The quality of attitude prediction was evaluated experimentally, working with four subjects. Subjects were between the ages of 18 and 28, and have kept diary-style weblogs for at least 2 years, with an average entry interval of three-to-four days. Digital personas were modeled from each subject’s weblog url.

In the interview, subjects and their corresponding generated models were asked to evaluate 12 short news snippets taken from Yahoo! News. The snippets are each approximately 150 words long, and 4 snippets were selected from each of three genres: social, business, and domestic. The same set of texts was presented to each participant and the examiner chose texts that were generally evocative. The subjects were asked to summarize their reaction by rating three factors on Likert-5 scales.

\[
V_{\text{spread}} = V_{\text{human}} - V_{\text{computer}}
\]  

We computed the mean spread and standard deviation across all episodes along each PAD dimension. On the –1.0 to +1.0 valence scale, the maximum spread is 2.0. Table 1 summarizes the results. Note that smaller spreads correspond to higher accuracy, and smaller standard deviation correspond to higher precision.

<table>
<thead>
<tr>
<th>Pleasure</th>
<th>Arousal</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean spread</td>
<td>std. dev.</td>
<td>mean spread</td>
</tr>
<tr>
<td>SUBJECT 1</td>
<td>0.39</td>
<td>0.38</td>
</tr>
<tr>
<td>SUBJECT 2</td>
<td>0.42</td>
<td>0.47</td>
</tr>
<tr>
<td>SUBJECT 3</td>
<td>0.22</td>
<td>0.21</td>
</tr>
<tr>
<td>SUBJECT 4</td>
<td>0.38</td>
<td>0.33</td>
</tr>
<tr>
<td>BASELINE$_{static}$</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>BASELINE$_{uniform}$</td>
<td>0.67</td>
<td></td>
</tr>
</tbody>
</table>

We give two baselines. BASELINE$_{static}$ always gives a neutral reaction, so the mean spread will be 0.50 on average. In the context of an interactive interface, BASELINE$_{static}$ is not a fair comparison because it would never produce any behavior. BASELINE$_{uniform}$ gives a random reaction from –1.0 to +1.0 assuming a uniform distribution, so the mean spread will be 0.67. The most realistic baseline would probably follow a Gaussian distribution, implying a mean spread whose lower bound is 0.50 and upper bound is 0.67.

On average, our approach performed noticeably better than both baselines, excelling particularly in predicting arousal, and having the most difficult predicting dominance. The standard deviations were very high, reflecting the observation that predictions were often either very close to the actual valence, or very far. The results along the arousal dimension recorded a mean spread of 0.22, and mean standard deviation of 0.20. This suggests that our attitude prediction models confidently outperform baselines in predicting arousal.

Table 2. Evaluating the effect of imprimers and LTEM.

<table>
<thead>
<tr>
<th></th>
<th>Pleasure mean spread</th>
<th>Arousal mean spread</th>
<th>Dominance mean spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imp ON, Epi ON (table 1 results sum’ed)</td>
<td>0.35</td>
<td>0.22</td>
<td>0.43</td>
</tr>
<tr>
<td>Imp ON, Epi OFF</td>
<td>0.34</td>
<td>0.21</td>
<td>0.43</td>
</tr>
<tr>
<td>Imp OFF, Epi ON</td>
<td>0.40</td>
<td>0.28</td>
<td>0.44</td>
</tr>
<tr>
<td>Imp OFF, Epi OFF</td>
<td>0.41</td>
<td>0.29</td>
<td>0.45</td>
</tr>
</tbody>
</table>

In the experiment, we also analyzed how often the episodic memory, reflexive memory, and imprimers were triggered. Episodes were on average, 4 sentences long. For each episode, reflexive memory was triggered an average of 21.5 times, episodic memory 0.8 times, and imprimers reflexive memory 4.2 times. To measure the effect of imprimers and episodic memories, we re-ran the experiment turning off imprimers only, episodic memory only, and both. Table 2 summarizes the results.

These results suggest that the positive effect of episodic memory was negligible on the results. This certainly has to do with its low rate of triggering. The texts being evaluated are news articles, but the episodic memories are formed from weblogs. This suggests that a linguistic incompatibility between the episodes in memory and the episodes been evaluated is to blame for low rates of triggering. A pleasant surprise in these admittedly preliminary results is that imprimers seem to play a measurable role in improving performance, which is a very promising result.

Overall, the evaluation demonstrates that the attitude prediction approach presented in this paper is promising, but needs further refinement. The highest accuracy and precision is demonstrated along the arousal dimension. The approach does quite well against the active BASELINE$_{uniform}$, putting it within the performance range of entertainment applications. But the alarmingly poor precision along the pleasure and dominance dimensions throw caution on other possible applications. Taking into account possible erroneous reactions, we were careful to pose What Would They Think? as a fail-soft interface. The reacting faces are evocative, and encourage the user to click on a face for further explanation. Used in this manner, the application is fail-soft because users can decide on the basis of the explanations whether the reaction is justified or mistaken. We do not suggest that the approach is yet ready for fail-hard applications, such as deployment as a sociable agent, because fallout (bad predictions) can be very costly in the realm of affective communication (Nass et al., 1994).

3.3 Evaluation of the Application
In addition to evaluating the underlying attitude prediction model, we also performed a user study to test the hypothesis that WWTT can help someone grasp the personalities and opinions of a panel of strangers more quickly and deeply than with baseline methods.

The subjects of the user study were 36 college students, formed into three comparable test groups. The examiner used WWTT to create a panel of four individuals who are extensive webloggers (at least two years of daily blogging), using their blogs as the
The study was posed as a “game” and the objective is for each subject to answer the most number of questions correctly regarding the general personalities and specific attitudes of the panel of four strangers, who are previously unknown to the subjects. They are allotted 20 minutes to answer 15 questions.

Subjects in test group #1 were allowed to read through the weblogs of the four individuals as their information resource. Test group #2 used a textual version of WWTT. Test group #3 used the real WWTT interface. The group #1 baseline represents how a user typically learns about the perspectives of people without the assistance of technology. The group #2 baseline provides a keyword-retrievable textual memory, controlling for all the non-affective elements of the application.

The textual version of WWTT warrants some further description. This version of the application does not use the color dimension (for expression of approval/disapproval) or the focus dimension (for expression of dominance/submission). Only arousal is expressed, through brightness. Whereas arousal is affective in the real WWTT, arousal in the textual version is proportional to the number of textual instances triggered by a textual episode. To put this another way, suppose a textual episode contained the concepts X and Y. X occurs 10 times in the personal texts, and Y occurs 19 times. Thus the total number of textual instances is 29, and the extent of the arousal reaction is proportional to this score.

In the explanation mechanism, the quotes are order ranked to make prominent quotes which overlap the maximum number of textual instances contained in the textual episode.

The community of personalities metaphor has been previously explored with Guides (Oren et al., 1990), a multi-character interface that assisted users in browsing a hypermedia database. Each guide embodied a specific character (e.g. preacher, miner, settler) with a unique “life story.” Presented with the current document that a user is browsing, each guide suggested a recommended follow-up document, motivated by the guide’s own point-of-view. Each guide’s recommendations were based on a manually constructed bag of “interests” keywords.

**Figure 4.** WWTT and two baselines in a person-learning task.

The 15 4-choice questions, given in random orders, fall into three categories of knowledge: general personality traits, specific attitudes explicitly contained in the weblogs, and specific attitudes not contained in the weblogs. The examiner is careful to ensure one clear answer for each question. The answers to the section on specific attitudes not contained in the weblogs were verified through consultation of the individuals depicted in the panel. The questions on general personality traits is of the vein, “who is the most shy?” Questions on specific attitudes explicitly contained in the weblogs are meant to test how well each of the three test groups can grasp existing attitudes, e.g.: “how does Sally feel about religion?” Questions on specific attitudes not contained in the weblogs are meant to test how well each of the groups can project how a panelist might react to something novel, e.g. “what would Sally think of Jim, given the bio on his web page?”

The accuracy of answers is summarized in Figure 4.

The results are exciting. WWTT consistently outperformed Control 2, came close to Control 1 in “personality traits,” and clearly outperformed Control 1 in “existing attitudes.” All three groups struggled with “novel attitudes” and performed comparably. Subjects in the Control 1 group commented that it was easy to build an overall picture of a person by skimming a more extended sample of their writing. However, search tasks required for “existing attitudes” handicapped Control 1, who had to use the search feature in the text editor, but could not query all four panelists in parallel as the WWTT interface enables. At first, subjects in Control 2 and the WWTT group struggled to come up with text to pose to the application. Many people in both groups came up with a surprisingly efficient strategy of passing in a string of keywords which would define the linguistic context probable to contain the information they wanted. For instance, to answer the question, “who loves to party the most?” a subject in the WWTT group typed something like, “party clubbing booze drinking drinks threw up” and then clicked on each face, reading salient quotes in the explanation to verify the attitude. When faced with a choice of which face to click first, subjects in the WWTT group usually clicked the one showing the highest arousal. Subjects in the WWTT group spent less time sifting through explanation quotes than subjects in Control 2, perhaps because affective saliency was a useful way to order quotes.

From this study, we can conclude that WWTT allows a user to more quickly and deeply grasp the personalities and specific attitudes of a panel of strangers than either of two baseline approaches. The results suggest that an affective memory can in many cases be a more useful way of organizing and presenting information than a purely textual memory. Despite the poor precision of attitude prediction as suggested in the evaluation of the underlying attitudes model, WWTT’s fail-soft explanation mechanism salvaged the usefulness of the attitudes prediction.

### 4. RELATED WORK

The community of personalities metaphor has been previously explored with Guides (Oren et al., 1990), a multi-character interface that assisted users in browsing a hypermedia database. Each guide embodied a specific character (e.g. preacher, miner, settler) with a unique “life story.” Presented with the current document that a user is browsing, each guide suggested a recommended follow-up document, motivated by the guide’s own point-of-view. Each guide’s recommendations were based on a manually constructed bag of “interests” keywords.

Our affective memory-based approach to modeling a person’s attitudes appears to be unique in the literature. Existing approaches to person modeling are of two kinds: behavior modeling, and demographic profiling. The former approach models the actions that users take within the context of an application domain. For example, intelligent tutoring systems track a person’s test performance (Sison & Shimura, 1998), while online bookstores track user purchasing and browsing habits and combine this with collaborative filtering to group similar users (Shardanand & Maes, 1995). The demographic approach uses gathered demo-
graphic information about a user to draw generalized conclusions about user preferences and behavior.

Neither of these two approaches are appropriate to the modeling of “digital personas.” In behavior modeling, knowledge of user action sequences are generally only meaningful in the context of a particular application and does not significantly contribute to a picture of a person’s attitudes and opinions. Demographic profiling tends to overgeneralize people by the categories they fit into, is not motivated by personal experience, and often requires additional user action such as filling out a user profile.

Memory-based modeling approaches have also been tried in related work on assistive agents. The Remembrance Agent (Rhodes & Starner, 1996) uses an associative memory to proactively suggest relevant information. Sunil Vemuri’s project, “What Was I Thinking?” (2004) is a mind prosthesis that records audio from a wearable device, and intelligently segments the audio into episodes, allowing the “audio memory” to be more easily browsed.

5. CONCLUSION

Understanding the perspectives of people we care about and having those perspectives available to us just-in-time during a task has been up to now a difficult problem with no good technological solutions. In this paper, we presented a novel intelligent user interface called “What Would They Think?” that observes what a user reads and writes and proactively shows the affective reactions of a configurable panel of people whose opinions are valued. Each person’s attitudes model is built automatically by mining attitudes from a corpus of personal text. A person-neutral affect sensing engine is adapted to the extraction of personal affect by averaging affective appraisals of the same text at various granularities, and by requiring stable attitudes to be conditioned over many exposures.

Both the underlying attitude prediction model and the application were evaluated in user studies. The results suggest that affective arousal is the most valuable of the PAD dimensions to making a person-centered evocative interface. It is also the arousal component of a person’s attitude that can be most accurately and precisely predicted using our approach. Using WWTT, study participants were able to more quickly and deeply grasp the personality and specific attitudes of a panel of strangers. We also learned that an affective memory is more helpful organization of a person’s attitudes than a purely textual memory.

The automated, memory-based personality modeling approach introduced in this paper represents a new direction in person modeling. Whereas behavior modeling only yields information about a person within some narrow application context, and whereas demographic profiling paints an overly generalized picture of a person and often requires a profile to be filled out, our modeling of a person’s attitudes from a “memory” of personal texts paints a richer, better-motivated picture about a person that has a wider range of potential applications than application-specific user models. Our user studies illustrate that the model for attitude prediction need not be perfect and free from erroneous predictions to usefully improve a user task. By offering an explanation mechanism, users can independently verify the validity of an affective reaction, and this lends all-important fail-softness to our interface.

In future work, we intend to give more prominence to the affective arousal dimension in the interface, as it is the component that can be most accurately predicted by the model. We would also like to investigate how a persona can be supported by the personas of other people, of social identities, etc. For example, a particularly strong belief such as “I love dogs” can cause the persona of the “dog-lover” identity to be attached to one’s own existing persona. Finally, we are also working on modeling personal attitudes from non-first-person texts, and investigating other applications for our person modeling approach, such as virtual mentors and guides, marketing, and document recommendation.

6. ACKNOWLEDGMENTS

The authors would like to thank Deb Roy, Roz Picard, Marvin Minsky, Cynthia Breazeal, Bruce Blumberg, Henry Lieberman, Ted Selker, and the blind reviewers for their feedback.

7. REFERENCES


