

humaine

D9a

Green paper

Kristina Höök and WP9 members



Date: 30 May 2005

IST project contract no.	507422
Project title	HUMAINE Human-Machine Interaction Network on Emotions
Contractual date of delivery	<i>May 31, 2005</i>
Actual date of delivery	<i>May 31, 2005</i>
Deliverable number	D9a
Deliverable title	Green paper
Type	Report
Number of pages	35
WP contributing to the deliverable	WP9
Task responsible	Kristina Höök
Author(s)	See page 5
EC Project Officer	Philippe Gelin

Address of lead author:
Kristina Höök
Department of Computer and Systems Sciences,
Forum 100,
164 40 Kista,
Sweden

Table of contents

1	THE PLACE OF THIS REPORT WITHIN HUMAINE	5
2	PROBLEM DESCRIPTIONS	6
2.1	TOWARDS A CONCURRENT MULTIMODAL EVALUATION	7
2.2	EVALUATING THE ROLE OF AFFECT IN ECAs AS SOCIAL ACTORS.....	8
2.3	UNDERSTANDING THE UTILITY OF ECAs IN AN APPLICATION AND THEIR ACCEPTANCE BY THE USER	10
2.4	ACCEPTANCE OF AN ECA BY THE USER AT LONG RUN	11
2.5	HOW TO CHOOSE THE BEST CHARACTER FOR CREATING EMOTIONS WITH THE USER	12
2.6	DATABASES OF ‘NATURALISTIC’ EMOTIONAL EXPRESSIONS: WHAT SHOULD THEY BE?	13
2.7	INDIVIDUAL AND SOCIAL EMOTIONS IN INTERACTION WITH TECHNOLOGY	15
2.8	COMPARISON OF ‘SUBJECTIVE’ AND ‘OBJECTIVE’ EVALUATION METHODS	16
2.9	MARK-UP LANGUAGE FOR ANALYZING RESULTS OF EVALUATION STUDIES	17
2.10	EVALUATION OF EMOTIONAL ARTEFACTS BY GROUPS OF USERS.....	18
2.11	CONTEXTUAL TASK-BASED MODELS OF EMOTION IN ACTION.....	19
2.12	BASELINE DATA ABOUT EMOTIONAL EXPRESSION	20
2.13	TEASING NARRATIVE AND EMOTION APART	21
2.14	EVALUATION OF EMOTIONAL DIALOGUE STRATEGIES IN PHONE-BASED SYSTEMS.....	22
3	REPORT FROM CHI-WORKSHOP.....	23
3.1	THE TWELVE PAPERS	23
4	CONCLUSIONS.....	31
5	REFERENCES	33

1 The place of this report within HUMAINE

The purpose of this green paper is to describe a range of problems that the researchers in WP9 of HUMAINE foresee as difficult usability and design issues that needs to be resolved in the future, and relate those to the WP9 exemplar. The purpose is not to provide an exhaustive list of all problems that exist, but to provide for some insight into how little has really been done in the area so far, and thus the great need for more both design and evaluation methods involving the end-users of affective interactive systems, criteria of success for these systems, and evidence for which kinds of application really make sense in the area.

Apart from the list of problems below, we also provide a report from a workshop at the CHI (Computer-Human Interaction) conference in 2005, organised by two HUMAINE members: Katherine Isbister and Kristina Höök (Isbister and Höök, 2005). The workshop was named “Evaluating Affective Interfaces: Innovative Approaches” and it attracted researchers spanning quite a huge range of theoretical perspectives on the problems, and thereby quite a huge span of possible methods. Again, it became clear that there are many white spots on the map and very little is understood in terms of getting the users feedback on designs or figuring out what measurements can be used to discuss success of systems. In addition researchers, designers and practitioners have their own views on what is important in a design process, something which further complicates matters. For researchers the most important thing is to *prove* the success of their systems using some suitable (usability)criteria. Designers on the other hand search for methods that support and provide *feedback to the design process*. Practitioners search for *good examples* that illustrate ideas in action and that can provide *practical advice* on how to build applications. In order to gain a complete picture of affective interaction design and evaluation, all of the above viewpoints need to be taken into account.

The range of suggestions for future research arising from this green paper should be seen as a small sample of all the research that needs to be performed in the area. Let us start with the list of problems handed in by researchers in HUMAINE and then we go on to summarising the workshop papers, before finally providing a short summary of how these problems relate to the WP9 exemplar and ongoing work in WP9.

The following persons have contributed to the deliverable:

Kristina Höök, Jarmo Laaksolahti, Fiorella de Rosis, Brigitte Krenn, Katherine Isbister, Benoit Morel, Asimina Vasalou, Catherine Pelachaud, Markus Ballegooy and WP9 members

The institutions that have contributed are:

KTH, OFAI, Imperial College of Science, Technology and Medicine, University of Paris 8, Università Degli Studi di Bari, T-Systems Nova GmbH, La Cantoche Production.

2 Problem descriptions

In order to collect information about important design and evaluation problems a document template was provided that WP9 members were asked to fill out. In addition to a description of the problem itself we asked for examples, i.e. situations when the problem was actually encountered or could be encountered, and suggestions about how to address the problem. Below is the returned set of problems that researchers in WP9 of HUMAINE foresee as difficult usability and design issues that needs to be resolved in the future. The list constitutes an inventory of the problems most important to WP9 researchers and is not intended to be exhaustive.

2.1 Towards a concurrent multimodal evaluation

Title/Summary:	Towards a concurrent multimodal evaluation
Authors:	Asimina Vasalou, Jeremy Pitt
Affected WPs:	N/A

2.1.1 Problem description

There is a need for emotion evaluation methods that capture the emotion as it occurs and evolves over time, which do not interfere during the emotion experience and that rely on multiple channels. We present several arguments to justify this claim, concluding with a solution to this problem.

It is known that spontaneous emotional expressions do not correspond to their later recollections resulting in a need for concurrent collection methods. Traditional post-assessment usability methods such as questionnaires and interviews do not suffice as they primarily collect post-experience data. To add to that, emotional experiences are not fleeting, but evolve over time. Only a time-based evaluation method which records the length and degradation of the emotion can be representative of the emotional experience. Moreover, as mentioned above, the method in place should be unobtrusive and not interfere during the user's interaction with a particular interface. To that effect, it has been demonstrated that techniques eliciting verbal *opinions* during task completion, decrease overall performance (e.g. Ericsson & Simon, 1984). And in consideration of the readings' accuracy, a given interference may interrupt the emotional experience so that its evolution over time is not captured. In light of this, the traditional prompting role of the facilitator should be reconsidered towards a less obtrusive 'elicitation' approach.

On account of the previous requirements, it is not surprising that a number of efforts have focused on inferring emotional states with the collection of corporal signals (i.e. physiological readings, facial expressions, vocal pitch and intonation). These channels indeed support the notion of time-based emotion capturing while interfering with the experience minimally. The weakness they pose however is the fact that they rely on inferences of a single channel. In certain cases, emotion may be expressed via one channel, while suppressed in all others. For example, when delivering a public speech, the speaker's heart rate may have culminated although his face appears calm and composed. This is further explicated by Ekman et al (1976) while investigating involuntary expressions during acts of deception. In attempting to deceive another, participants were able to control their facial expressions but not their vocal pitch.

2.1.2 Suggestions for how the problem may be addressed

In conclusion, we intend to address two issues in WP9. First, we aim towards developing a concurrent, time based and unobtrusive evaluation method. Our particular interest is focused on an extension of the cognitive think aloud protocol. Second, we intend to test the validity of readings derived by a single channel (e.g. vocal) versus multiple channels towards determining the validity of a multi-channelled approach.

2.2 Evaluating the role of affect in ECAs as social actors

Title/Summary:	Evaluating the role of affect in ECAs as social actors
Authors:	Brigitte Krenn
Affected WPs:	WP9, WP6, WP3

2.2.1 Problem description

There is a broad and constantly increasing variety of work on evaluating ECAs. (See for instance Ruttkay & Pelachaud (2004) for a recent compilation of approaches.) This variety of loose ends reflects the fact that ECAs are a fairly novel technology which is intended to make the human-computer interface more human-like and thus more natural, intuitive, appealing, trustworthy, etc., with all these attributes themselves addressing rather vague concepts. Secondly, there is the whole complex of “humans being inherently social”, and thus are inclined to treat computers as social actors (cf. Nass & Moon 2000), and even more so to treat ECAs as social actors. And last but not least, there is the whole bundle of questions relating to modelling affect in ECAs.

2.2.2 Suggestions for how the problem may be addressed

Give suggestions for how one might address the problem or at least take steps towards solving it. This can include work needed by other work packages or even other projects.

In order to tie up some of these loose ends, we need to first of all make the following basic assumptions.

To begin with, we need to distinguish different aspects of affect. In particular we will start out from Scherer’s classification of “affective states”, i.e., emotion, mood, interpersonal stance, attitude, and personality traits (cf. Scherer 2000).

We assume that not all affective states are equally relevant to be modelled for the ECA on the one hand, and the user model at the other hand. For instance, interpersonal stances and attitudes will be more important at the ECA side, whereas a model of the current user emotion may be highly desirable in a variety of applications, especially if we assume that the user’s emotion has strong impact on how the human-machine interaction proceeds, and what the user (consciously and unconsciously) draws from the interaction. In this respect, the central question is by means of which cues ECAs are able to influence the emotions, and in the long run the mood and the attitude of the user towards the system. Thus the focus needs to be put on the question by which means does the ECA *induce* a certain emotion in the user and NOT how emotion can be represented in the ECA.

Another assumption we make is that ECAs are perceived by humans as social actors, i.e., interaction with ECAs leads humans to start to establish social relationships with the ECA. For a recent discussion of arguments for social effects of ECAs see Kraemer (2004). See also evidence from the “Dice experiment” by Rehm and Andre comprising two human players and an ECA. Thus evaluating affective ECAs is always also evaluating the ECA as a *perceived* social actor in certain setting.

[The view of ECAs as social actors closely relates to the work of Nicole Kraemer, Dept. of Psychology, University of Cologne. Input from her work would be very welcome.]

As regards experimentation, a potential scenario could be to make use of the SALAS metaphor of listening ECAs and create small, well defined WoZ-type settings for user-ECA interaction where not only affective display but also social aspects of the ECA-user relation are

taken into account. A crucial but not yet examined question at all is what happens in a more long term ECA-user interaction.

2.3 Understanding the utility of ECAs in an application and their acceptance by the user

Title/Summary:	Understanding the utility of ECAs in an application and their acceptance by the user
Authors:	Catherine Pelachaud
Affected WPs:	WP6 and WP9 might want to have a dialogue about the problem

2.3.1 Problem description

Several ECAs systems have been developed so far. Evaluation studies have been performed to determine if an ECA is of any use for a given application. Often the result of these studies is ‘that depends’. It depends on the role the ECA is given in an application (banker or teacher); it depends if the user has to trust the ECA or not; it depends on the seriousness of the conversation topic (eating disorder or shop seller); etc. Another dimension of consideration is on what consists the added-value the ECA brings: aids for teaching (measure of the learning curve), entertainment (measure of the time spent with the agent), task efficiency (measure of the overall time a user took to perform a task), having fun... The utility of an ECA may also depend on how the ECA is accepted by users. Studies have shown that users prefer ECAs that look like them (in term of cultural roots, personality,...), that are expressive rather than neutral. ECAs may understand user’s affective states as well as display affective behaviors. For some applications ECAs can trigger or modulate user’s emotion. A capability of pedagogical agents is often to be able to motivate users, to cheer them, to detect negative state of users, etc. But ECA may have negative drawback: they are difficult to understand due to the quality of their animation and of the speech synthesizer; they are not believable; they are far from being really interactive. It might take longer and more constraining to view an animation than to read a text that describes a set of instruction. Their presence may become very annoying after a while. So many factors enter in the evaluation of the usefulness and acceptance of ECAs in applications that the scalability of any evaluation result to other systems seems difficult.

2.3.2 Suggestions for how the problem may be addressed

Developers of ECA systems and experts in evaluation studies could work together a methodology that, given an application to build, to elaborate a set of specifications of ECAs. The problem would be tackled from one aspect: the emotion of the user and of the ECA.

2.4 Acceptance of an ECA by the user at long run

Title/Summary:	Acceptance of an ECA by the user at long run
Authors:	Catherine Pelachaud
Affected WPs:	WP6 and WP9 might want to have a dialogue about the problem

2.4.1 Problem description

When ECA systems are being evaluated, it is often based on interaction with users that last a very short time. But how can we infer from these studies how an ECA is perceived in the long run in an application? An ECA may be fun to interact with at the beginning but its pervasiveness may be extremely annoying after a while. Should an ECA develop a companion relationship with the user or should it remains the perfect silent butler? If a relationship is built between a user and an ECA, it evolves through time. How far should it go? ECA systems are developing techniques to learn about the user's knowledge, habits, emotional states, etc. How much personal information do a user accept an ECA knows and shows it knows it?

One aspect of this problem is linked to the type of behaviors the ECAs exhibit. Which behavioral style an ECA should display: exaggerated, simplistic behavior, few behavior, emotional behavior...? How should an ECA behave when it has nothing to say or it listens to the user; how should it bring and maintain the attention of the user; should an ECA behave naturalistically or overdo it...

2.4.2 Suggestions for how the problem may be addressed

If we address mainly the behavioral aspect, an idea is to use the WoZ technique. A WoZ study could involve users interacting with ECAs behaving with different expressive styles. An animation system allowing displaying different styles of ECA's behaviours could be developed. It could be based on existing tools. It requires developing an appropriate interface that allows the wizard to control and modify ECA behaviour on-line. A main difficulty is to derive such an interface: simple control mechanisms have to be developed to allow the wizard to control the ECA in real-time.

2.5 How to choose the best character for creating emotions with the user

Title/Summary:	How to choose the best character for creating emotions with the user
Authors:	Benoît Morel
Affected WPs:	WP6 and WP8

2.5.1 Problem description

One of the main problems using ECA is to select and design the character. What is the best character to create, depending of the target (culture, age, gender,..), the media support (PC, mobile phone, SetTopBox,..) and the mission (companion, assistant,..). Very often we realize that a Human character is not adapted. Sometimes, an animal is more convenient or even an object. In these cases, designers humanize the graphical representation by adding eyes, mouth, arms, hands, feet, in order to show some expressions and create emotions. Because of the selected design, some emotions are harder to show or easier.

We realize that for the same goal, different representations (human, animal, object,..) can be done and it's always hard to decide.

2.5.2 Examples

An important French commerce company asks us to design a character for their web site. This company sells DIY objects and the mission of the character is to orient the visitor and to "keep" him a long time on the web site. When he discovers the virtual character, the visitor needs to understand immediately where he is, what kind of tools he can find on this web site and of course the visitor has to be seduced by the character. The first impression is always the most important and so the design of the character is really important to catch the best emotion. In this case, several designs can be done as human (for example, a DIY man) or humanize objects (A DIY box).

In order to design this character, we spend a lot of time with our client (the company) in order to know more about the company, the products, the visitors, the mission, and the needs. Then we suggest different designs trying to show a general expression on the character in order to help our client to make a decision.

2.5.3 Suggestions for how the problem may be addressed

It will be interesting to create a generic survey for people who want to decide which ECA they want for their use and thanks to this survey, orient the designers.

2.6 Databases of ‘naturalistic’ emotional expressions: what should they be?

Title/Summary:	Databases of ‘naturalistic’ emotional expressions: what should they be?
Authors:	Kristina Höök
Affected WPs:	WP5 and WP9 might want to have a dialogue about the problem

2.6.1 Problem description

When designing an ECA or a device for recognising emotions, databases with naturalistic emotional expressions is key in the design process. They are used both for modelling of expressions and input patterns, but also later for verifying the resulting systems. The aim of HUMAINE is find ways by which the gathering and annotation of data enables gathering of such naturalistic data – a problem that should not be underestimated.

However, the design process for an affective application does not only entail specific frozen moments of interaction when an ECA needs to display a particular facial expression or the signs/signals system infers a specific emotional state. When designing for an affective interaction, the interaction unfolds over several interaction steps and emotions arise from an active act of interpretation and participation from the end-user side. In order for the databases to be an inspirational source for this part of the design process, they need to show how a cue from one person elicits an emotional cue from the other, or how the context in one setting creates the basis for the next situation. Suchman has criticized the cognitivist reductionist basis for the field of affective computing, reducing human emotional responses to discrete, universal component parts, arising from some underlying state, and offers examples of alternative designs that put emotional interaction at the core but where: “we might differently conceptualise affective encounters at the interface: as irreducibly contingent meetings of particularly situated persons with equally particular, dynamic, and culturally inflected things.”

Thus, the problem also lies in creating databases that are not devoid of contextual data or claims to be relevant to all sorts of interactions, but where more of the context is made part of the database description.

2.6.2 Examples

An early experiment in WP9 focused on gathering input for the design process of a “gossiping agents”. The idea was to create a simple leisure application for the mobile phone where the user would be able to interact with a character that gossips about movie stars and exhibits various extreme emotional expressions. To create such a character, we would need to know more about how to portray extreme emotional facial and bodily expressions, but perhaps more importantly, we would need to know the timings of jokes (a joke will fall flatly if told too slowly or too fast), how to create contradictions between emotions portrayed through facial expressions and what is said in order to create humours irony, floor-grabbing techniques for when there is a good pause in the interaction for coming up with the next topic, and similar.

We collected data from three different meetings where 4 – 6 women met to discuss movie stars and gossip about them. We found that we need to annotate these video snippets with markings showing a number of situations more to do with how the interaction unfolds rather than specific emotional expressions or states, as for example:

- imitation is really funny and a powerful tool that people use

- making oneself the butt of the joke by having an unpopular opinion or a ‘dumb thought’ (often together with submissive body language)
- pre-cuing others of how to feel by imitating the opposite

2.6.3 Suggestions for how the problem may be addressed

Collaboration between designers of systems and database creators might be interesting in order to see what the differences in needs and goals consists of.

2.7 Individual and social emotions in interaction with technology

Title/Summary:	Individual and social emotions in interaction with technology
Authors:	Fiorella de Rosis
Affected WPs:	WP9

2.7.1 Problem description

- In interaction with technology, emotions of various kinds may be established: they *may be* ‘individual’ and in this case are much lighter than the ‘classical’ set of emotions in the OCC classification (anxiety, relief, satisfaction). But they may also be ‘social’ (irritation, sympathy or antipathy, tenderness, contempt, sense of belonging, ...).
- The first category is mainly related to the difficulty in performing tasks or in their outcome. The second category results from the attitude of users towards the application and is particularly relevant in those interaction forms which involve some kind of antropomorphization of the technology, such as those with ECAs. The question is then: do the two categories need the same method to be recognized, modelled, interpreted? probably not. And (immediately related to the previous ones): which methods may be employed to evaluate interaction from this viewpoint?
- In particular:
 - the concept of ‘social emotion’ is tightly associated with the concept of ‘empathy’, although they are not identical, and the exact form of social emotion which might be established in interaction with technology merits to be discussed and deepened.
 - which ‘natural’ methods may be employed, to enable users to express their -light individual or social- emotions?
 - how to recognize these emotion from natural language analysis of the user moves, from speech, from facial expressions, from gestures etc?
 - which methods to employ to iteratively design systems which respond to the social relation needs of their users?

2.7.2 Examples

Almost all those who worked recently on ECAs considered the problem of social relationship with these agents, by using different terms and applying different theories. Recently, the problem was discussed in a Meeting in Vienna that was organized by OEFAL. Workshops on Empathy have been organized by some HUMAINE participants.

2.7.3 Suggestions for how the problem may be addressed

We might promote reflection on the concepts of ‘social interaction’, ‘social emotion’, ‘empathy’ and related ones within the Network. This might be seen as a subtask of WP3 or as a subject of discussion in one of the HUMAINE Workshops. At least the following Humaine Partners might be involved: Isabella Poggi, Ana Paiva, Anton Nijholt and Brigitte Krenn to discuss the concepts of ‘empathy’ and ‘social emotion’ with Kia Höök for ‘immediate’ and ‘natural’ methods to enable users to express this kind of emotions

2.8 Comparison of ‘subjective’ and ‘objective’ evaluation methods

Title/Summary:	Comparison of ‘subjective’ and ‘objective’ evaluation methods
Authors:	Fiorella de Rosis
Affected WPs:	WP9

2.8.1 Problem description

Artefacts have been evaluated since long-time by collecting (in controlled studies by means of questionnaires of various kinds) the users’ subjective feeling towards the application and its interaction modes. This method has been applied, as well, to evaluate the emotional reaction of users to these artefacts, although the questionnaire items were, in this case, less easy to define and the way users may be asked to express their perceived emotional state was the object of some debate. More recently, other, more ‘natural and immediate’ forms of subjective evaluation have been proposed (like clicking on icons, selecting a colour and so on).

By ‘objective’ evaluation methods, we mean those methods which are based on observation of the user behaviour when interacting with the artefact, on tracing this behaviour by several means and on analyzing these data to extract some quantitative evaluation parameters. Data may range from natural language texts to gestures etc. Some examples of parameters which may be extracted from this kind of data and apply to dialogs:

- degree of involvement in the application
- level of initiative of the user in the interaction
- intensity and valence of the emotional state
- ... and similar

Both kinds of evaluation criteria are probably useful, and it is likely that they, on one side, enable evaluating different aspects of interaction and, on the other side, are in a way - positively or negatively- correlated. The purpose of this task is to compare the two categories of methods to find out the advantages and disadvantages of each, and their possible correlations.

2.8.2 Examples

We applied both kinds of metrics to evaluate the user attitude in a health promotion dialog and could verify that the relationship between subjective and objective evaluation was not as clear as one would expect, and not in the most ‘obvious’ direction. In this case, we measured ‘objective’ data from linguistic analysis of the user moves in the dialog and ‘subjective’ data by coupling ‘natural’ interaction (clicking on icons expressing ‘positive’ vs ‘negative’ vs ‘unclear’ concepts) with a final questionnaire. Objective measures were obtained by quantitative and qualitative analysis of the user contributions to the dialog.

2.8.3 Suggestions for how the problem may be addressed

Apparently, this is a subject of interest to the community of Evaluation (see the Workshop at CHI): it might be interesting to introduce it as one of the subjects of the next WP9-HUMAINE Workshop, to check which the experiences within the Network are.

2.9 *Mark-up language for analyzing results of evaluation studies*

Title/Summary:	Mark-up language for analyzing results of evaluation studies
Authors:	Fiorella de Rosis
Affected WPs:	WP9, with WP5 and WP6

2.9.1 Problem description

Mark-up languages are the object of investigation in various workpackages of HUMAINE, for various purposes:

In WP5, they play the role of labelling the corpus of data collected in various conditions and various forms, and aim at providing a background of how emotions may manifest themselves in various situations.

In WP6, (where they are called ‘reference languages’) they aim at defining a standard for formulating input to Embodied Animated agents, by describing the signs of emotion that these agents should be able to show.

If results of evaluation studies performed in different contexts and by different groups have to be quantified, they also need a common reference language. It is likely that this language has its own needs and is different from the two languages mentioned before. It is influenced on one side by the experimental asset which was defined for the evaluation study and on the other side by the kind of parameters the experimenters believe are meaningful to connote the user behaviour.

Defining a mark-up language also involves defining a method for labelling data collected by a group of raters, of indices to measure the level of agreement among raters and of criteria for defining when agreement reaches a level that justifies accepting a given result. Also in this case, results of evaluation methods have their own problems which merit to be discussed and contrasted with the problems found in the two contexts mentioned before.

2.9.2 Examples

Discussion on how to label corpora according to emotion (valence, intensity and emotion names or categories) may be found in a large number of papers. The choice of the language is closely related to the concept of emotion which is applied. Fewer references may be found on how to label corpora according to ‘social emotions’ or ‘attitudes’.

2.9.3 Suggestions for how the problem may be addressed

This is typically a cross-WP topic: the groups working on Reference Languages, on Databases and on Evaluation might meet to discuss the solutions they are elaborating and to propose an agreed solution. A contribution of WP3 might be essential, to produce a proposal which clearly shows the interdependence between the concept of affect which is considered and the markup language defined.

2.10 Evaluation of emotional artefacts by groups of users

Title/Summary:	Evaluation of emotional artefacts by groups of users
Authors:	Fiorella de Rosis
Affected WPs:	WP9 with WP8

2.10.1 Problem description

Public applications, which address themselves to groups of users rather than to individual ones, are becoming frequent, especially with new interaction paradigms like ubiquitous computing.

In this kind of applications, several users will be affectively influenced, to different degrees, by the application: but they will also influence each other to a different extent and with different degrees, depending on their personality, the role they take in the group and so on.

Evaluating the emotional impact of an application, in this kind of situations, requires specific methods. Individual reactions cannot just summed up to reveal the overall reaction of the group: on the contrary, the dynamics of these reactions, the way contagion is established, the way individual reactions evolve after contagion should be investigated.

Measurement methods may combine subjective with objective evaluation methods, but probably the second method will be more meaningful, and new methods of data collection and analysis will have to be defined.

2.10.2 Examples

Advertising is a typical example. All cases of group participation in health promotion or self-care provide good examples of collective vs individual persuasion and evaluation processes.

2.10.3 Suggestions for how the problem may be addressed

The subject of formulating persuasion strategies addressed to groups of users seems to be of interest of WP8. We might combine those who are interested in this particular domain in WP8 and WP9, to strengthen the efforts. Considering the limited experience acquired in this domain by HUMAINE participants, some integration with external competences might be advisable.

2.11 Contextual task-based models of emotion in action

Title/Summary:	Contextual task-based models of emotion in action
Authors:	Katherine Isbister
Affected WPs:	From WP3 for WP 9 and perhaps also WP6

2.11.1 Problem description

In attempting to evaluate peoples' emotions in response to systems, WP9 members have realized that some emotions arising in the course of human-computer interaction do not necessarily neatly fit existing models of emotion as they are described in the literature we have learned about. Instead of extremes of disgust or happiness or fear, or other socially expressive emotions, users tend to report they feel milder, task-based emotions such as contentment, frustration, confusion, and the feeling of flow or disruption of flow. These emotions occur within the context of the user's expectations of the system and his/her goals for using the system.

We would like to be able to apply psychological theory to the study of these emotions as they arise in interaction contexts with systems, and we are hoping that our colleagues in WP3 can provide some guidance along these lines.

2.11.2 Examples

When a person is working with an application (such as an internet browser) and things are going smoothly, s/he may describe being in a state of flow. This flow may be interrupted by the interface doing something unexpected, creating surprise and perhaps frustration or confusion. These are the kinds of emotions we would like to address with our evaluation tools—perhaps being able to predict when they should occur and then measuring them, as well as understanding ways to measure them.

2.11.3 Suggestions for how the problem may be addressed

The theory group (WP3) could let us know about any existing task-based models, and could work with WP9 and WP6 to discuss the types of emotions that tend to come up in interaction and that we wish to evaluate. In addition, we could do a literature survey of existing emotion assessment in the context of interface tasks (such as the work of Picard's group at MIT) and bring this to a discussion with WP3 to help flush out relevant literature and theories.

2.12 Baseline data about emotional expression

Title/Summary:	Baseline data about emotional expression
Authors:	Katherine Isbister
Affected WPs:	From WP4 to WP9 (and also relevant to WP6?)

2.12.1 Problem description

In order to evaluate whether system users are experiencing an emotion, we need to better understand the signals of various emotions, in particular emotions that are relevant to the context at hand. We would like to have a taxonomy of signals of emotion, with baseline data that can be used to compare user responses with.

2.12.2 Examples

Knowing the signals of frustration, for example, helps the Affective Computing group at MIT to choose detection mechanisms as well as to recognize when the emotion is occurring in response to a system. We would like to see a summary of these signals and baseline data to use as guidelines depending upon input modality (e.g. video imagery, skin conductance, etc).

2.12.3 Suggestions for how the problem may be addressed

The signs and signals group (WP4) could collate what is known, and provide this to our group to aid in our evaluation process. This data might also be used to experiment with interaction design with emotional interfaces (e.g. WP6)

2.13 Teasing narrative and emotion apart

Title/Summary:	Teasing narrative and emotion apart.
Authors:	Jarmo Laaksolahti, Kristina Höök
Affected WPs:	WP8, WP6

2.13.1 Problem description

It has been argued that establishing a narrative context is important for making emotions meaningful and understandable (see e.g., Rizzo 1999) For instance *a situation or interaction history* is necessary for characters in games to be understandable (Sengers, 1999, 2003). A player's interaction with – and experience of – a character becomes meaningful and rich once something is known about the character's background (back-story), its drives and motivations and it is placed in a narrative context. At the same time emotions are an important aspect of character believability (see e.g. Bates, 1994) and expressiveness which also affects how users perceive narratives. Hence developers creating applications that aim to provide both a narrative and emotional experience face the problem of teasing narrative and emotion apart. Although a user's final experience is a combination of both, it may be desirable to tease the two apart to understand where to focus efforts during development. We may for instance have a good story but characters affective cues may not be salient enough to make an impact on users, or the affective cues may be just right but as the scenario is lacking a good story they are meaningless. Given the complex relationships between narrative and emotion there are a number of unresolved issues. Taking games as an example: first it is unclear how to evaluate users' affective experiences of a game, second it is unclear how to evaluate users' narrative experiences of a game and finally it is unclear to which extent narrative context and emotions *can* be separated. If emotional and narrative experience can indeed be separated which should be developed/evaluated first? Or maybe there is some way of doing it in parallel in a structured fashion?

2.13.2 Suggestions for how the problem may be addressed

A literature survey of how professionals within the arts evaluate their work (especially film and literature) could constitute a first step. Then we could look at (or construct) cross disciplinary approaches involving methods from both the arts (film, literature, etc) and the sciences that attempts to capture narrative and affective experiences simultaneously. Another approach would be to use a tiered evaluation model where first one aspect of the application is evaluated and then the other.

2.14 Evaluation of emotional dialogue strategies in Phone-based Systems

Title/Summary:	Evaluation of emotional dialogue strategies in Phone-based Systems.
Authors:	Markus van Ballegooy T-Systems
Affected WPs:	WP 9

2.14.1 Problem description

In the last years, some emotional and phone based systems have been developed that are explicitly designed to affect the user's emotional state. Especially when the caller is classified as being angry or the communication situation seems to get critical, the system applies certain dialogue strategies or communication tactics that will lead the communication to a "happy end". Those strategies are usually designed by using human communication behaviour as a model. Nevertheless, the effectiveness of those tactics is still to be proved. It seems to be difficult to determine valid indicators for the strategies' "effectiveness" or success.

2.14.2 Examples

We set up a self service application where customers can look up their telephone bills and find information about their mobile phone tariffs.

Within dialog situations where slight anger was detected and where there were no further hints that the users request could not be handled within the voice portal, we used a dialogue strategy that is called "conciliation by mirroring". The goal of this strategy is to show the user that his emotions are recognized but that it is better to continue the task. Immediately after an utterance that was classified as angry by the acoustic classifier, our system interrupted the main dialog by playing a short prompt like: "I see that you are a little bit excited. So it will be the best if I continue quickly with your query!"

Within dialog situations where strong anger was detected in combination with hints that the dialog will not be completed successfully we used a strategy called "conciliation by empathy and delegation" by offering the user the possibility to leave the voice portal and speak to a "real" service agent. Before the connection was put through, the system verbally showed empathy for the user's situation. We used a prompt like: "I notice that you are angry because I don't grasp your request. I can really understand you! The best thing will be if I put you through to a service agent."

2.14.3 Suggestions for how the problem may be addressed

We think the problem of indicators should be addressed by using several methods. We are about to design a pilot study, where the system is implemented in a commercial self service application.

We could run a quasi – experiments, where one user group is exposed to the emotional communication tactics and another group will use the "normal" system without emotional behaviour (the kind of strategy used could also be a second experimental factor).

Dependent variables could be subjective estimations of the emotional effect caused by the strategy. In addition, objective measures like completion rates, mood changes should be used to cross-validate the subjective measures.

3 Report from CHI-workshop

Two HUMAINE members, Katherine Isbister and Kristina Höök, organised a workshop on evaluation of affective interfaces at the CHI (Computer-Human Interaction) conference in Portland, Oregon (Isbister and Höök, 2005). The workshop was named “Evaluating Affective Interfaces: Innovative Approaches” and 12 papers were selected for presentation at the workshop – each shortly summarised below. The researchers at the workshop presented work that spanned a large range of theoretical perspectives and thereby quite a large span of possible methods for evaluation. In our perspective, this shows the diversity of what is meant by affective interactive systems and also how different researchers view what can be seen as successful design.

There are several reasons why the organisers of the workshop felt that the CHI conference was the best place to do this workshop. There is strengthening and continued interest within the CHI community in designing affective engagement with interfaces. Affect is an important part of user engagement with games, interactive narrative, synthetic characters and robots, wearables, voice interfaces, and many other interactive systems. Systems designed to promote community or to enhance safety, two key themes at the 2005 year’s conference, also benefit from consideration of users’ emotional states. Core CHI practitioners have promoted the value of thoughtfully crafting emotional qualities of interfaces (e.g. Don Norman’s 2003 CHI keynote; October issue of *Interactions* featuring ‘Funology’). Research has advanced our abilities to detect affect in users and incorporate this into system response (e.g. Picard, 2000), and there is a steady trickle of papers presented at CHI and other ACM conferences on affectively engaging systems.

As of yet, however, there has not been adequate discussion within the CHI community of how the evaluation of such systems presents unique challenges to researchers and industry practitioners. How do we measure whether an end user has been affected emotionally in the way we’d like, by a system? Do we look for physiological evidence of emotions? Do we ask for a self-report of emotional state? If the latter, will traditional questionnaire methods elicit accurate responses? Emotional responses could be distorted and disturbed by traditional HCI methods such as ‘think aloud’, and settings such as the average usability laboratory. CHI community members and practitioners outside the field have been experimenting with tactics for gathering good data about users’ emotional reactions toward systems in order to improve design. The workshop provided a forum for discussing these efforts.

Another key question is how and when we engage users in the design and evaluation cycle, in order to create truly engaging affective interactions. What sorts of prototypes are best for testing an affective interface concept? What kind of user participation produces the best results? New methods and strategies may be called for, as well as the borrowing and modifying of tactics from other fields, such as film, advertising, and consumer product design.

The workshop goals were therefore to bring together those who have been exploring and innovating in the affective interface evaluation domain, in order to:

- Bring together examples of affective interface evaluation strategies already in use.
- Put together a list of current best practices, and collect a body of references from past efforts to evaluate affective reactions to designed systems (both successes and failures), to help us all leverage what is already known.
- Identify key challenges and issues for future work.

3.1 *The twelve papers*

The workshop attracted a substantial set of papers out of which twelve were chosen. The papers themselves can be found at:

http://www.sics.se/~kia/evaluating_affective_interfaces/papers.html

But let us briefly summarise each here to give a perspective on how different researchers address what we think is one of the main problems that HUMAINE WP9 should tackle in its exemplar (see deliverables D9b and D9c): finding methods to evaluate affective interaction applications in such a way that it provides feedback to the design and can further the research field.

3.1.1 I like it - Affective Control of Information Flow in a Personalized Mobile Museum Guide

The first presentation came from a group which is active in HUMAINE (Goren-Bar et al., 2005). They evaluated their museum guide in a user study in order to see whether allowing museum visitors to express their affective appraisal of different parts of the tour through the museum, could create a better personalisation of the route through the museum and in general, a better experience of the museum. The belief behind their museum guide was that an ‘optimal’ multimedia tourist guide should support strong personalization of all the information provided in a museum in an effort to ensure that each visitor be allowed to accommodate and interpret the visit according to his own pace and interests. In such a guide, an interaction based on expressing affective attitude may improve efficacy of the interface in particular when, like in museums, the technology should not hinder the “real” experience.

Based on the results from their user study, they were able to improve their interface for affective feedback from the users. Through a *like-o-meter*, the visitors can provide their feedback on the different frescos in the Buonconsiglio Castle museum, see Figure 1.



Figure 1 The latest PEACH interface

The paper presented their study and how it had been able to provide them with substantial information on the mistakes in their first design. Thus, this paper presents an example of how a user study can provide valuable design feedback.

3.1.2 Evaluating a Fabric Device Controller

The second presentation at the workshop, (Hurst and Zimmerman, 2005) was aiming at a quite different kind of affective system. Their goal was to create a device controller (for television and other kinds of equipment in the home) that could be used by elderly users. The goal was to find a design that was enjoyable, aesthetically pleasing, and that could be integrated with a chair. Enjoyability is an important emotion to measure because it often influ-

ences the long-term relationship users have with a product and the companies that provides the product.

The system that Hurst and Zimmerman had constructed, was a novel device controller that that was ‘organically integrated’ into the armrest of a lounge chair, where it is meant to be used. In addition, they had developed a study to measure the enjoyability users experience with the armrest device controller compared to two other device controllers. At the time of writing the paper, they had not yet performed the study, but they had some interesting ideas on how to measure enjoyability. Their proposal was four dependant measures: *personality, usability, engagement, and emotional response*.

The personality measure would be done through categorizing the participants into interest types by administering part of the Strong Vocational Interest Blank (SVIB). This test is a vocational scale that has participants rank activities, occupations and amusements according to whether they like, dislike or are indifferent to them. In addition, personality would be measured through an *Openness to New Experience* scale to learn more about their personality.

To the authors, usability is an important measure because they: “doubted that participants will enjoy interacting with a device that they cannot use”. The usability would be measured in a traditional sense through measuring efficiency with the device.

For the engagement variables, they would apply the eight dimensions of enjoyment (Bartneck, 2003) and measuring them through allowing subjects to interact with the device freely when they feel that they are unsupervised.

Finally, for the enjoyability measurement, they aimed to create an enjoyability questionnaire, using Likert scale statements.

While this paper discussed future work and tentative ways of seeing evaluation, it was interesting because it was addressing a physical, textural design, rather than the typical interactive characters we see in HUMAINE. It also had some interesting perspectives on which measurements are important – not only usability and efficiency, but also enjoyability, personality, etc.

3.1.3 Affective computing vs. usability? Insights of using traditional usability evaluation methods

Wiberg presented her thesis work where she had tried traditional usability methods, such as heuristic evaluation by Jacob Nielsen (1993), and extended them somewhat to see if they could be applied to the problem of evaluating websites aiming to be *fun*. The results show that the methods are applicable but need revision:

“When it comes to the development of inspection methods, the challenges include finding proper heuristics to support the experts in using Heuristic Evaluation, providing conditions for experts which bridge the gap between evaluation and authentic use, developing complementary methods for use in combination with existing methods etc. In empirical evaluation of entertainment in the context of web usability, the most crucial aspect might be to consider how to arrange a setting that is as natural and authentic as possible when evaluating fun, as this seems to be important for the results. Overall, the results of the study clearly show that important aspects of affective interfaces can be revealed by using traditional usability evaluation methods – aspects which should be considered early in the design phase.” (Wiberg, 2005).

The important lessons learnt here for HUMAINE purposes, is to not disregard traditional usability methods when we move to affective interaction. This is particularly interesting for “cheap” methods such as heuristic evaluation that can be performed without having to rely on lots and lots of end-users. Such methods have been shown to substantially improve the design and its usability in a cost-efficient way.

3.1.4 AMUSE: A tool for evaluating affective interfaces

Noel Chateau and Marc Mersiol presented their tool AMUSE, that is aimed at supporting user evaluation of affective interfaces in general, even if their example in the paper focused on ECAs (Embodied Conversational Agents). Their tool allows for logging of the system-user interaction together with the user's eye-gaze and physiological data. There are several important problems with these kinds of loggings: synchronisation of data logs from the different sources is key, as well as how to provide statistics that summarises the results in meaningful ways.

Through using their logging tool on an example application they were able to provide findings such as:

“By crossing eye-gaze and electro-physiological data, AMUSE allows to analyse the values of the mean of the skin conductance signals when the eye gaze are in a given zone of the screen. Such an analysis was conducted for the text and the ECA zones of the screen, for the three voice conditions, and showed that there were no significant differences between zones and between voices ($F(2,24)=0,99$ $p=0,39$). Therefore, although voice quality seems to have an effect on users' eye gaze (driving user's attention on the screen), it seems to have no effect on their emotional reaction as measured by skin conductance.” (Chateau and Mersiol, 2005).

The authors themselves point out that these kinds of physiological measurements needs to be complemented by posteriori questionnaires to find the subjective measurements, and to provide a complete picture of the effects of an affective system.

3.1.5 Methodologies for evaluating the affective experience of a mediated interaction

Cahour et al. (2005) presented a method based on a phenomenological theoretical perspective. Thus in contrast to Wiberg, Chateau and Mersiol, and Goren-bar et al., this is not at all based on a traditional cognitivistic stance, but on meaning-making and reflection. Cahour et al. tries to get at the *experiential* aspects of an interface – in some ways similar to Hurst and Zimmerman.

While usability is addressing important issues, the methods and criteria from that field are still not able to get at the factors that make people bring technology into their lives and use it for real. The main criteria of evaluation are performance and productivity in reference to a so-called 'cognitive efficiency'. In this respect, the importance of *meaning* from the user's point of view of a non-effective or unexpected action is crucial to better understand the rationales behind a particular behaviour.

This is why the authors propose to make use of methods that better address people's own accounts of the meaning they are experiencing from some particular IT-artefact. The author propose to use methods such as an “explicitation interview”:

“It is the aim of the "explicitation interview" to help the interviewee to render explicit what was only implicit in her description, or even implicitly present in her experience; it implies some form of becoming aware, of explicit apprehension of a content that was present in the experience but not yet apprehended, and as such implicit for the subject himself.” (Cahour et al., 2005)

Obviously, getting at the meaning of an artefact or its experiential properties is a very hard goal to achieve. Thus, the paper goes through an example study that the authors have performed and shows how this process can be done. These kinds of approaches are probably quite alien to most HUMAINE researchers who come from a more science-and-engineering perspective. Despite this, it is our belief that from the perspective of actually addressing users' experiences of the systems we aim to create, these kinds of methods will probably be more important, if not key to the development of successful affective interactive systems. The prob-

lem at hand really lies in understanding the experiences users have with our systems on this pre-reflected, experiential level.

3.1.6 Evaluating Affective Computing Environments Using Physiological Measures

The previous paper stands in stark contrast to the paper presented by Mandryk. Her approach does not address any of the semantic, meaning-making processes of the users, but instead only the physiological aspects of users' experiences. This is obviously a very important aspect of users' experiences of systems, since we know that emotions are experienced not only by our conscious thinking, but also throughout the whole body. Whether this is the *true* account of our emotional experience is another story.

Still, Mandryk has a point when she says that:

“Current subjective methods of evaluating entertainment technology aren't sufficiently robust. Our research project aims to test the efficacy of physiological measures as evaluators of collaborative user experience with play technologies. We found evidence that there is a different physiological response in the body when playing against a computer versus playing against a friend. These physiological results are mirrored in the subjective reports provided by the participants. This research provides an initial step towards using physiological responses to objectively evaluate a user's experience with collaborative play technology.” (Mandryk, 2005)

Mandryk has tested her measurement technology on hockey games and similar. Mandryk's aim is to complement the “objective” measurements she can collect through physiological responses with subjective measurements.

3.1.7 Dealing with User Experience and Affective Evaluation in HCI Design: A Repertory Grid Approach

Fällman and Waterworth presented a very interesting method for capturing the experiential and subjective aspects of affective designs without resorting to questionnaires and other simplistic ways of getting at subjective experiences. The problem with subjective experiences is that either the subjects are allowed to express themselves freely rendering the evaluator with lots and lots of qualitative data that cannot be structured and measured across subjects, or the evaluators will set the boundaries of what can be said through deciding on the questions in a questionnaire or similar. With the repertory grid approach (RGT) there is a way around this problem.

RGT has its root in Kelly's *Personal Construct Theory* (Kelly, 1955):

“While it is not necessary to fully buy into the underlying theory to use RGT in practice, Kelly argued that we make sense of our world through our own ‘construing’ of it. That is, we tend to model what we find in the world according to a number of personal constructs, which are bipolar in nature”.

According to Kelly, we judge for instance other people through forming construct such as Tall—Short, Light— Heavy, Handsome—Ugly, and so on. A ‘construct’ is hence a single dimension of meaning for a person allowing two phenomena to be seen as similar and thereby as different from a third [...]. Kelly suggested RGT as a structured procedure for eliciting a repertoire of these conceptual constructs and for investigating and exploring their structure and interrelations [...].” (Fällman and Waterworth, 2005).

From this basis, the method is constructed. Subjects are asked to classify three objects at a time, telling the experimenter which ones belong together and which does not. They are then asked to tell the experimenter why the two belong together and not the third. This classification brings out a set of bi-polar scales that are then used to describe the qualities of the designed artefact at hand. The data from all users is entered into a gigantic grid with all the sub-

jective scales that all the subjects have come up with. Through collapsing those that belong together, patterns of descriptions arise. Thus, the method is both quantitative and qualitative – and most important, it is built on how users experience the objects – not on pre-conceived notions of what the qualities are or should be put together by the researchers.

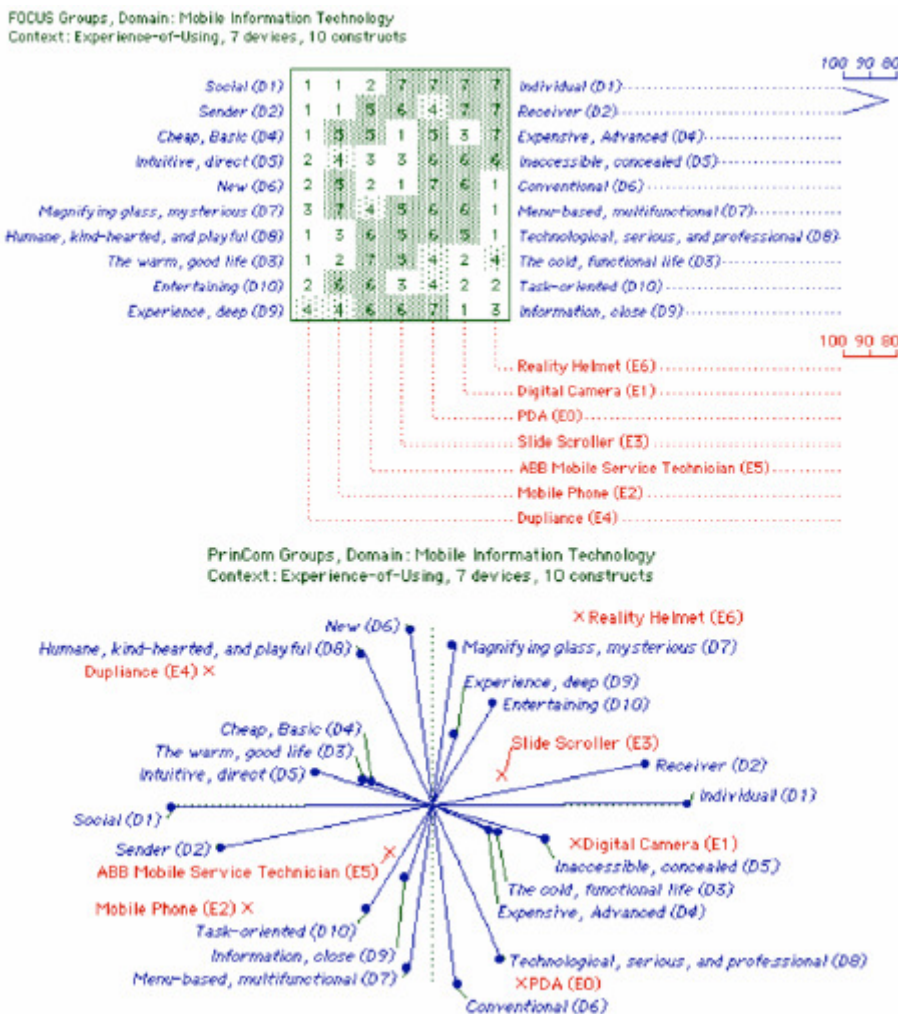


Figure 2. The ten resulting dimensions from one of the experiments in Fällman and Waterworth’s experiments.

3.1.8 Insight Into Strong Emotional Experiences Through Memory

A similar point about the subjective experiences of affective interactive systems was made by Mentis (2005). Her main point was that what we remember after interacting with some system, even as late as 2 weeks after interacting with it, might be a more important clue to what works in a system and what does not than many other measurements. This is, of course, most relevant to those kinds of systems that are designed to create strong emotional experiences. Mentis points out that this is also a potentially very cheap way of measuring affective involvement. While many other methods involves expensive equipment for measuring eye-gaze and similar, a method asking for what users remember after interacting with a system can be much cheaper.

Mentis also provides a range of concrete advice on how to perform interviews aiming at extracting what is remembered. For example, one should not outright ask subjects about their emotions unless you are interested in some particular emotion. Instead very open-ended questions work better: “tell me about your experience with this system”.

3.1.9 Evaluating affective interactions: Alternatives to asking what users feel

Picard and Bryant Daily provides a very good overview of a range of behaviours that can be observed and what is known to date about how well they work. With this list, Picard and Bryant Daily wants to argue for taking a set of physiological signals and body movements as input to an analysis of whether a system is inducing the kind of affective response intended. In particular, they want to place those measurements relative to the tasks that users are performing in order to really understand what the measurement mean in relationship to the context of use.

In their conclusions, they write:

“We have highlighted several means of assessing affect beyond directly asking somebody what they are feeling. One can imagine an interface interaction in which the user’s facial and electro-dermal activity are monitored for valence and arousal information, and the interaction is followed by an assessment task such as the Duncker’s Candle Test, where better success is expected with more positive affect. Thus, information about the user’s affect can be gleaned from the physiological sources as well as the task performance. This data could additionally be compared against self-reported measures. There is a lot of room for new methods to be discovered; the ones we have presented here are just a few of the possibilities.” (Picard and Bryant Daily, 2005).

The set of methods and ways of observing behaviours presented in this paper is probably of high interest to many of the HUMAINE researchers.

3.1.10 Sensual Evaluation Instrument

At the workshop, there was also a presentation directly based on work within WP9 by Höök, Isbister and Laaksolahti: the *sensual evaluation instrument* was presented. More information on this idea can be found in previous deliverables from this WP (D9b, D9c). The main addition was that we were able to present the designed objects to be used in the evaluation method at this workshop. This will be reported in full detail in the deliverables that will follow in WP9 later during the project.

3.1.11 Intimate Objects: A site for affective evaluation

In the paper by Kaye, we were presented with a minimal design for contacting your beloved one. Through clicking on a Virtual Intimate Object (VIO) in the taskbar of your PC, the corresponding “dot” on your partner’s taskbar would change colour (see Figure 3).



Figure 3 Virtual Intimate Object (VIO) in taskbar, showing color changes over a twelve hour period. Note rapid initial fading in top line. Final image shows display of remote partner’s button state on mouseover.

The problem presented in the paper lies in how to evaluate this kind of intimate technology to be used between partners in a couple. Kaye proposes to make use of a design method for evaluation here: the cultural probes by Gaver and colleagues (Gaver et. al, 1999). In particular, Kaye makes use of a logbook that the users themselves fill in. He also asks a number of provocative questions that he aims to be fun to fill in for the subjects. The reason is of course that otherwise, people will not be very willing to participate in a study that reveals quite some intimate details in their patterns for contacting their partners. The subjective data is complemented by some objective log data on how often the system is used and by whom.

3.1.12 Evaluating Affecter: Co-Interpreting What Work

Finally, the last paper was presented by Sengers et al. (2005). It followed the same tradition, in a sense, as the papers by Fällman and Waterworth, Cahour et al., and Kaye, in that it attempts to address the meaning-making aspects of interaction between human and machine. It employs a phenomenological perspective on evaluation and use.

The authors have created an affective system they name *Affecter*. In *Affecter*, a still picture of the user is transmitted from the user's office to someone else's office. The picture is slightly distorted in ways that the user decides upon herself/himself. These distortions are used to convey the user's emotional state. The interpretation of the distortions is not given, but is negotiated between the two users as part of their usage process over time.

To evaluate such a system, the authors propose to use a combination of three methods. One is a critical practice evaluation:

“The evaluation of design and design processes, which we refer to here as critical practice evaluation, is a typical approach in critical studies and the humanities. Here, the guiding question is why something was designed as it was in the first place. The choices made by the designers of the system are primarily under review; the goal is to identify, examine, and possibly challenge the cultural, social, and historical influences at play in the design process.” (Sengers et al., 2005).

This critical practice evaluation is combined with a way of looking upon systems as an appropriation process: users pick up technology and use it in ways that makes sense to them – not always in the ways anticipated by the designers. Finally, they also attempt to integrate system evaluation into the process. The three are not necessarily compatible, and thus the approach is highly experimental.

In the evaluation of *Affecter*, Sengers et al. propose that the designer herself will take part in the evaluation, gradually evolving an understanding of how this piece of equipment is used in her life together with her next-door office friend. This is a quite unusual method, pulling the subjective stance to an extreme. This is probably quite provocative to most researchers in HUMAINE, but seen from a humanities viewpoint, not a very strange stance to take.

4 Conclusions

The problems presented above spans a wide range of areas and contexts, and pose different requirements on the type of research that is required to address them. What is important however is that they all fit within the WP9 exemplar *A Framework for Design and Evaluation of Usable Affective Interaction Applications*. The framework entails an assortment of techniques and methods for various aspects of the design process including generating initial ideas for affective applications, refining ideas into systems/products or validating the affective interaction loop in finished systems/products. As a whole the framework is intended to be a broad resource for designing and evaluating affective systems – representing first and foremost a user-centred perspective on affective interaction applications. The tools and methods of the framework will be applied and evaluated in a variety of situations/applications/domains. Records from the sessions will provide valuable guidance for future users of the tools and methods regarding their proper usage and expected results. The exemplar consists of four main elements:

1. *Criteria for usable affective interaction systems*. Develop an understanding for what makes affective interaction systems successful and formulate some criteria. These criteria will not be objective, independently measurable entities, but will make sense relative to the specific application domain, aim to capture subjective experiences of the application, and foremost, be related to the designer's intention for the application.
2. *Evaluation metrics for criteria*. Translate criteria for affective interaction systems into some metrics and suggested methods.
3. *Existing user-centred methods for design and evaluation*. Condense experiences from applying existing user-centred design methods to the design and evaluation of affective interaction systems – which methods work and which do not? Examples of such methods include body-storming (Oulasvirta et al, 2003), interaction re-labelling (Djajaningrat, et al., 2000) and personas (Cooper, 1999).
4. *New methods for design and evaluation*. Develop new methods for capturing unique aspects of affective interaction that can be used during an iterative design-evaluate-redesign process. Methods that will be investigated include: a sensual method for non-verbal mediation of affective state, a Wizard-of-Oz environment for multimodal emotional interaction, and an extended think-aloud protocol designed to capture emotional interactions.

Table 1 summarizes the relation between problems stated by HUMAINE researchers in this document and the exemplar elements in which they are predominantly addressed. As can be seen all exemplar elements address at least one problem. Furthermore the different elements of the exemplar appeal to the researcher, designer and practitioner perspectives (as discussed in the introduction) to various degrees. Usability criteria and evaluation metrics are more research oriented whereas *existing evaluation methods* and especially *new evaluation methods* also accommodate design and practitioner perspectives. All in all the exemplar covers all three perspectives fairly well.

The *existing evaluation methods* element addresses fewer problems than the other elements. This is however to be expected as evaluation of affective interaction is something traditional

methods were not meant to cope with. It also reflects the amount of work that goes into the element by HUMAINE partners. Nevertheless there are situations when using an existing method, or a slightly altered versions thereof, is appropriate. Outside of HUMAINE Wiberg (2005) has for instance investigated the use of traditional usability methods for measuring “fun” in applications. She concludes that existing methods can be used sometimes, but usually require some amount of modification to be useful.

The problem descriptions also suggest important topics for the upcoming WP9 workshop to be held in early 2006. For instance *usability criteria* and *evaluation metrics* recur in almost all problem descriptions. Evaluation of affective interaction is relatively new so criteria and metrics have yet to emerge. Furthermore affective interaction is highly application and context dependent. Hence what is needed is perhaps not a collection of ready made criteria and metrics but instead methods for creating them for each unique situation.

Problem	Usability criteria	Evaluation metrics	Existing evaluation methods	New evaluation methods
2.1				X
2.2	X	X		
2.3	X			
2.4		X	X	
2.5	X			
2.6	X	X		
2.7		X	X	X
2.8	X	X		
2.9		X		X
2.10	X	X		X
2.11		X		
2.12		X		
2.13	X	X		X
2.14	X	X	X	

Table 1 Relation between exemplar elements and problems

In general, design and evaluation problems gleaned from the CHI Workshop are also addressed by one or more of the exemplar elements although not explicitly stated here. This indicates that WP9 is conducting innovative research that is of interest to the research community within the area and is well positioned to take a leading role in the definition of affective interaction design and evaluation.

5 References

- Bartneck, C., Interacting with an embodied emotional character. In Proceedings of the 2003 international conference on Designing pleasurable products and interfaces, ACM Press, 55-60. (2003).
- Bates, J. 1994. The role of emotion in believable agents. Communications of the ACM, 37(7):122 - 125.
- Cooper, A. (1999). *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Re-store the Sanity*, SAMS, 1999.
- Craggs, R., Annotating emotion in dialogue: issues and approaches. In M. Lee (Ed), Proceedings of the 7th Annual CLUK Research Colloquium, 2004.
- Craggs, R., McGee Wood, M., A two-dimensional annotation scheme for emotion in dialogue. Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: theories and applications. 2004.
- Djajadiningrat, J. P., Gaver, W. W., and Fres, J. W. (2000) Interaction relabelling and extreme characters: methods for exploring aesthetic interactions, Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques, Pages: 66 – 71, 2000, ACM Press New York, NY, USA
- Ekman, P., Friesen, W. V., & Scherer, K. R. (1976). Body movement and voice pitch in deceptive interaction. *Semiotica*, 16, 23-27.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: Bradford Books/MIT Press.
- Gaver, B., Dunne, T., Pacenti, E. Cultural Probes. *Interactions*, 6(1) 21-29 (1999).
- Isbister, K, and Höök, K. (2005) Evaluating affective interfaces: innovative approaches, CHI '05 extended abstracts on Human factors in computing systems, Portland, OR, USA, Pages: 2119 – 2119, ACM Press. (Workshop proceedings can be found at: http://www.sics.se/~kia/evaluating_affective_interfaces/)
- Kelly, G. *The psychology of personal constructs*. Vol 1 & 2. Routledge, London, UK, 1955.
- Kraemer N. (2004). Social effects of embodied conversational agents. International Workshop on Dimensions of Sociality: Shaping Relationships with Machines, Vienna, 18.-20. November 2004.
- Nielsen, J. (1993). *Usability Engineering*. Academic Press.

Oulasvirta, A., Kurvinen, E., and Kankainen, T. (2003) Understanding contexts by being there: case studies in bodystorming Personal and Ubiquitous Computing, Volume 7, Issue 2 (July 2003) Pages: 125 – 134, Springer-Verlag London, UK.

Rizzo, P., (1999) *Emotional Agents for User Entertainment: Discussing the Underlying Assumptions*. International Workshop on Affect in Interactions: Towards a New Generation of Interfaces, Sienna, Italy.

Ruttkay Z., Pelachaud C. (eds.), *From Brows to Trust Evaluating Embodied Conversational Agents*, Kluwer Academic Publishers, 2004.

Scherer, K. R. (2000). Psychological models of emotion. In J. Borod (Ed.). *The neuropsychology of emotion* (pp. 137-166). Oxford/New York: Oxford University Press.

Sengers, P. 1999. Narrative Intelligence. In Dautenhahn, K., editor, *HumanCognition and Social Agent Technology*, volume 19, pp. 1 - 26. John Benjamins Publishing Company.

Sengers, P., Schizophrenia and narrative in artificial agents. In Mateas, M., Sengers, P., editors, *Narrative Intelligence*, pp. 259-278, John Benjamins Publishing Company.

Picard, R. 2000. Perceptual user interfaces: Affective perception. *Communications of the ACM*, 43 (3). *Interactions* 2004. Funology issue, September/October

Workshop on “Evaluating Affective Interfaces: Innovative Approaches” at CHI 2005

Isbister, K, and Höök, K. (2005) Evaluating affective interfaces: innovative approaches, CHI '05 extended abstracts on Human factors in computing systems, Portland, OR, USA, Pages: 2119 – 2119, ACM Press. (Workshop proceedings can be found at: http://www.sics.se/~kia/evaluating_affective_interfaces/)

B. Cahour, P. Salembier, Ch. Brassac, J.L. Bouraoui, B. Pachoud, P. Vermersch, M. Zouinar: Methodologies for evaluating the affective experience of a mediated interaction

Chateau, N., Mersiö, M.: AMUSE: A tool for evaluating affective interfaces

Fällman, D., Waterworth, J.: Dealing with User Experience and Affective Evaluation in HCI Design: A Repertory Grid Approach

Goren-Bar, D., Graziola, I., Pianesi, F., Rocchi, C., Stock, O., Zancanaro, M.: I like it - Affective Control of Information Flow in a Personalized Mobile Museum Guide

Hurst, A., Zimmerman, J.: Evaluating a Fabric Device Controller

Höök, K., Isbister, K., Laakso, J.: Sensual Evaluation Instrument

Kaye, Joseph Jofish: Intimate Objects: A site for affective evaluation

Mandryk, R. L.: Evaluating Affective Computing Environments Using Physiological Measures

Mentis, H. M.: Insight Into Strong Emotional Experiences Through Memory

Picard, R. W., Daily, S. B.: Evaluating affective interactions: Alternatives to asking what users feel

Sengers, P., Boehner, K., Warner, S., Jenkins, T: Evaluating Affecter: Co-Interpreting What Work

Wiberg, C.: Usability and Fun

Wiberg, C.: Affective computing vs. usability? Insights of using traditional usability evaluation methods

HUMAINE deliverables

Deliverable D9b: Preliminary Plans for Exemplars – Usability. Available at <http://emotion-research.net/deliverables/D9b.pdf>

Deliverable D9c: Potential exemplars – Usability. Available at <http://emotion-research.net/deliverables/D9c%20potential%20exemplars%20usability.pdf>