Chapter 5

USER-CENTRED DESIGN AND EVALUATION OF AFFECTIVE INTERFACES

A Two-tiered Model

Kristina Höök

> *What a piece of work is a man! How noble in reason! How infinite in faculty! In form and moving, how express and admirable!*
>
> —Shakespeare, Hamlet

Abstract One obvious challenge for affective interfaces is to find ways of checking whether the expressed emotions are understood by users, and whether the system can interpret user emotions correctly. Even more challenging is whether the overall usage scenarios are achieving their purpose of being e.g. engaging, fun, believable, or creating a relationship with the user, and how much of this can be attributed to the emotion modeling and expression. We propose a two-tiered design and evaluation model. We exemplify this model through studies of three different affective interfaces: the *Agneta & Frida* system, the *Influencing Machine*, and *SenToy & FantasyA*.

Keywords: Affective interaction, synthetic characters, user evaluation, user-centred design.

Z. Ruttkay and C. Pelachaud (eds.), From Brows to Trust, 127–160.

1. Introduction

Affective computing, or the development of computational systems which can be aware of and respond to human emotions, has become the focus of a great deal of attention in the Artificial Intelligence (AI) community.

Recent developments, such as the results from Tristão & Isolda (Martinho and Paiva (1999)), and Influencing Machine (Sengers et al. (2002), Höök et al. (2003)) suggest that a too narrow understanding of emotions will fail to address the important issues in interaction. The AI-approaches to affective computing often focus on what one might call an 'informatics of affect', in which emotions are treated as units of information. Emotions are analyzed, classified, discretized, and formulated as units whose purpose is to inform cognition or be communicated. The often-used integrative cognitive theory of emotion of Ortony, Collins and Clore (1988), for example, defines emotions in terms of a set of discrete, basic types and focuses on the cognition or reasoning which may give rise to them. Once a set of emotional units is defined, input devices can be designed which can turn physiological responses into information. For example, Fernandez, Scheirer, and Picard's (1999) Expression Glasses measure the movement of facial muscles and classify the resulting expression into a small, discrete set of emotions. Ark, Dryer and Lu's (1999) Emotion Mouse extends a normal computer mouse to deduce users' emotional states from physiological information such as pulse and galvanic skin response.

Frequently in this tradition, emotions are subsumed to rationality or effectiveness. Damásio's (1994) influential arguments for the importance of emotion in scientific research, for example, gain currency from the idea that emotion is necessary for true rational behavior. Similarly, Picard's (1997) ground-breaking work on Affective Computing argues that computers must be able to process emotion in order to function maximally effectively with human beings.

While defining, classifying, creating logical structure for, and understanding the relationship of rationality to emotions can be useful exercises, we believe this mindset is in danger of missing a fundamental point: affect is not just a formal, computational construct, but also a human, rich, complex, and ill-defined *experience*. Rationalizing it may be necessary to make it computable, but an affective computation that truly inspires and incorporates human emotion must include a broader cultural perspective, in which the elusive and non-rational character of emotion does not need to be explained away (Sengers et al. (2002)). From this perspective, computation may be used, not to acquire and reason about user's emotional states, but rather to create intuitive experiences of affect by the user during interaction.

A substantial design challenge in constructing a technical system that creates intuitive experiences and supports open interpretation, then, is the need to bridge the rational objectivity of the software and the hardware with the interpretational complexity of users' subjective experiences. Doing this well requires insights into how to develop the design. The line of argument presented here is that design and evaluation methods placing the user and usage at core, can be one key component in achieving the design goals of affective applications. Our starting point is a set of user studies the author has performed previously that we shall revisit and to some extent reanalyze. The methods that we have found to be most useful in capturing the idea of user experience are open-ended, subjective, interpretative studies performed through a two-tiered method. The first step in this method is to get the interface expression and interpretation right (usability). The second, more interesting step is to try and evaluate whether the affective aspects of the system do indeed contribute to the overall goal of the system, and users' experiences.

The work presented in this chapter should therefore be seen as an attempt to show that user studies interwoven into the design process can be crucial in the design process, but only if we can move away from simplistic measurements that 'prove' the efficiency of our affective interactive systems, and instead aim at deeper, interpretative understandings of what is really going on between user and system.

Let us start by outlining our philosophy underlying our method and the specifics of the method. We shall then go through previous work and in particular turn to a set of user studies performed according to our ideas[1]: a study of the Agneta & Frida system (Höök et al. (2000)), two studies of the Influencing Machine (Sengers et al. (2002), Höök et al. (2003)), and two studies of SenToy and FantasyA (Andersson et al. (2002), Paiva et al. (2003), Höök et al. (2003)). While all three systems are aimed at invoking affective responses from the user, they also examplify three quite different forms of Embodied Conversational Agents (ECAs), which is the focus of this book.

2. Underlying Philosophy and Method

As indicated above, the prevailing approach in the design of affective interaction is to construct an individual cognitive model of affect from first principles, implement it in a system that attempts to recognize users' emotional states through measuring biosignals, and through this try to achieve an as life-like or human-like interaction as possible, seamlessly adapting to the user's emotional state and influencing it through the use of characters in the interface or other affective expressions.

There has been quite some research on how to recognize users' emotional states through singular, one-off, readings of biosensor data, facial expressions, body posture, interaction with devices, such as mouse or keypad, or props, such as plush toys. However, repeatedly there seems to be the same conclusion: while some basic emotions (fear, stress, and arousal) may be recognized, the methods fail to get the whole picture and often contradictory results arise between users' self-reports of what they think and feel and their physical expressions (e.g Höök et al. (2000)). They also fail to understand any more complex and interesting emotional states that users might be in – such as shame, guilt, positive arousal, or flow.

It is probably impossible to detect fine-grained aspects of human emotion. People are interesting intelligent beings, and their emotion processing does not constitute some simple stimulus/response model. Human emotion relates to so many complex interactions that no modeling will ever be able to "detect" them. We have a personality, a mood, attitudes and value systems that are individual as well as cognitively related, we have bodily states that we influence and bodily states that we cannot influence (hormone levels, diseases...); we are influenced by the current context, and so on. An emotion state is usually not a single state – it is a mixture of several emotions along several scales such as arousal level, energy involved, long-lasting moods, more cognitively-induced versus more bodily-induced emotions, or valence (positive/negative) of the emotion. You might be in a melancholic mood that lies like a blanket on top of any emotion you have, or you might be in a context that does not allow for jumping around and thereby experiencing and reinforcing the strength of the inner emotion. It is hard to envision any modeling system that would be able to deal with and mimic such a complex and changing situation. As one of the studies discussed below showed, facial expressions of users only reveal one tiny aspects of how and why users react in certain ways to affective systems. Personality, value systems, ethics, and other individual differences also come into play as determinants to why we react in certain ways to these systems.

But the problem we would like to discuss here is not the problem of understanding how complex the human mind is, or how difficult it will be to try and correctly recognize users' emotional states from simple measurements of facial expressions or other biosignals, since we would like to stay clear of discussing counterarguments such as that this could be described as a problem due to lack of knowledge of how to model human emotions in machines, lack of sensors to recognise emotion states, or lack of correct theory of the human emotional processing system and consequently lack of good, computational models of emotions that can be inserted into these affective systems. Instead, we would like to argue that what is more crucial in creating affective interactive systems, is to understand how to influence the users' emotional states and be able to maintain and build user emotional involvement to create a coherent cognitive and emotional experience. With such a goal, bad modeling of human emotions lacking respect for the complexity of our inner life can be devastating. On the other hand, rightly used, affective interaction based on some emotional models can make us learn more, make better decisions, understand each other better in social applications and shared workspaces, and sometimes simply enjoy the application more.

Creating such systems is, obviously, a hard and very difficult goal to achieve. We know for sure that movies, novels, television shows, arts and music are indeed able to get people affectively involved. But we want to make end-users affectively touched by interacting with systems that model emotions, reason using emotions and express emotions. How can we aid the design process and make it more likely that we succeed? Our argument here is that one tool in the repertoire, among many others, could be a user-centered development methods. A user-centered approach throughout the development of affective interactive systems will aid designers to at least stay on track, focused on the end-user experience, even if it does not provide the whole answer to how to design these systems.

2.1 Our Philosophy

Our approach in the design of affective interaction has therefore had another starting point than that taken in Affective Computing. Our user-centred perspective does, in turn, influence how we think user evaluation studies should be done. We base our work on the following three assertions:

> **Assertion 1:** *People's affective reactions are parts of ongoing interactions embedded in a broader social context.*

People's affective interaction consists of much more than what can be understood from simplistic local measurements of their bodily reactions. Significant emotions (beyond elemental experiences such as of surprise, disorientation, or disgust) are to a large extent social phenomena that take place in specific cultural settings, taking on particular expressions colored by the culture and the group of people at a particular place. The meaning and expression of emotions like guilt or shame are given both by their local social context as well as by their cultural context.

Assertion 2: *Affective interaction has a broader scope*

Affective outputs should not be seen as an end product but rather be made part of the interactive coupling. Through affective input through affective toys, tangible input media, or affectiveware, and acquisition of an understanding of the affective output these generate, users will become more or less part of the system and will be more or less affectively involved. We believe that it is crucial to tap into those affective input and output modalities that speak more directly to our affective states, such as soundscapes, colors, imagery, and tactile media.

Assertion 3: *People's affective reactions are adapted to the current context*

Through experience, by watching others, by studying the specific culture at places, people will learn to portray affect through different behaviors under different circumstances. Thus, someone might scream out loud in happiness at a soccer game but only smirk in a research project meeting, all because of context and interaction with others and the setting. This becomes particularly relevant when we invent novel ways for users to interact affectively. While we can be inspired by theories of human emotion, the particular interactions we invent have to be designed and developed in ways that are particular to a specific activity and its purpose. We must be aware that people will pick up and learn how to interact in ways that are given by the specific interaction devices, the context and purpose of use, and the expressive behavior of the system. This interaction cycle has to be developed in a user-participatory design cycle in order to identify the particular difficulties and opportunities for design.

2.2 Our Method

There are very few user studies of the short-term and even fewer of the long-term effects of affective interaction. On the other hand, designers of artifacts, artists, musicians, writers, people in advertising, and more recently web- and game designers have played around with evoking emotions for ages. What differs here is the *interaction* between the artifact aimed at raising emotions or expressing emotions and the viewers'/listeners'/readers'/users' reactions and (affective) actions at the interface. Users will be involved in the loop in a more active manner – expressing their own emotions rather than only be influenced.

A lot of the work on affective interfaces is focused on implementing affective interaction through interactive characters, but affective interaction may also be realized in various other ways. In many affective interaction scenarios (besides interactive characters), the goal is to entertain. The HCI community has only recently started to debate how to take those characteristics into account when performing usability studies or providing input to design. These aspects are sometimes referred to as *hedonic usability factors* (Hassenzahl et al. (2000)) or pleasure-based human factors. Affective interfaces may also, of course, be used as part of learning systems, e-commerce applications, or general desk-top applications.

Open Interpretation Since the field of affective interaction is fairly new, there is no general agreement on what to evaluate through a user study. Researchers in the field have been focused on issues like natural expressions, perfect models of the user's emotions, design of sensors and readings of sensor-data, and not really concerned with whether this aim for naturalness or the emotion models as such, do in fact contribute to the overall success of the system. While Bates, who first coined the expression *believability* of characters[2], was aiming for a design that could suspend disbelief (1994), other researchers have been using the concept believability in the more simplistic sense of 'naturalness' of face, body and voice of characters. The idea of 'suspension of disbelief' as coined by Disney, has been misinterpreted as meaning as 'human-like as possible'. As put by Persson et al. (2002) when discussing how to create Socially-Intelligent Agents (SIA):

> In order to develop believable SIAs we do not have to know how beliefs-desires and intentions *actually* relate to each other in the real minds of real people. If we want to create the impression of an artificial social agent driven by beliefs and desires, it is enough to draw on investigations on how people in different cultures develop and use theories of mind to understand the behaviors of others. SIAs need to model the *folk-theory reasoning*, not the real thing. To a shallow AI approach, a model of mind based on folk-psychology is as valid as one based on cognitive theory.

The approach suggested by Persson et al. is to look upon human-computer interaction (the 'shallow AI approach') as a constructivist perspective on users where they themselves make sense and create meaning out of their interactions with the world. Thus, instead of viewing end-users as passive viewers of what the 'perfect' system is constructing based on models of their emotional states, end-users are viewed as active co-constructers of meaning. Our approach is to agree with this perspective and add some practical methods for understanding how users react to the kinds of systems we want to build in order to further the understanding of the design process.

Informal Methods Studies in other fields, such as natural language interfaces, adaptive interfaces and intelligent user interfaces show that there are principles and peculiarities particular to the design of human machine interaction (Dahlbäck et al. (1993), Höök, (2000)). A computer system is a designed artifact – not a 'natural' thing. While the field of HCI certainly recognizes that there are design considerations that should be built from knowledge of human abilities and limitations (see e.g., Norman (1990)), they also recognize that computers are part of human culture, and thus subject to change. Over and over, artifacts are designed that users then take into use in ways that are quite different from what the designer expected (Suchman (1987)). A design process that fails to involve end-users in the design loop, will fail to recognize the particular quirks and problems of how to design these artifacts.

Within HCI, formal user studies (quantitative-scientific) are the gold standard for evaluating computational systems. But the aim in the affective interaction systems might not be best captured using formal user studies as these rarely are able to capture end-user experience (in a broader sense). We believe that informality and open-ended interpretation of users experience is key here as done in the more ethnographically inspired parts of HCI. This approach is similar to how artwork is evaluated through art critics and informal encounters between the artist and the audience. This will not render results that are independent of time and culture – but the point is that no user evaluation studies are independent of time and culture anyway[3], something that we come back to below.

Informality can, e.g., be observed in the HCI literature on evaluation of art-influenced speculative design. For example, the Presence project was evaluated informally by describing the designers' experience in installing the system and observing user interaction (Gaver et al. (2001)).

Anecdotal evidence, informal chats between users and system-builders, tiny study sizes, forms structured to influence user interpretation, no discussion or analysis of results: this may sound like a to-do list for bad evaluation. But since the goal is to aid the process of improving the design until the end-user experience and the system interaction harmonize, we prefer a rich, narrative, and singular understanding before a simpler but rigorous and generalizable understanding (Höök et al. (2003)). This interest in singularity and narrative complexity allies well with the recent ethnographic turn in HCI; yet many ethnographers may feel uncomfortable in promulgating a personal vision to users to the same extent as we have done in some of the studies discussed below.

No Averaging – No Normal User In looking for this rich, narrative, constructive understanding of what is going on between user and system, we are not looking for the average user reaction. We are interested in the richness and complexity of unique, individual users, cultural contexts, and resulting variety of interpretations and experiences of the system. Since affective interactive systems in many cases will make end-users engage in complex acts of interpretation, it would not be appropriate to summarize the results of a study into a few statements that are said to hold for everyone. Also, the statistical averaging and laboratory simplifications necessary for reliable scientific statements may wash out all the details that interest us.

Thus, we are not looking for representative user groups, or generalisable scientific results that last for ever – we are looking for input to the design process.

Two-tiered Method In our experience from the user studies and design work with the three systems presented here, we noted that it was necessary to divide the user studies into two different levels. The first obvious challenge for affective interfaces is to find ways of checking whether the expressed emotions are understood by users, and whether the system can interpret user emotions correctly. It might be that a design of an affective interactive character is perfectly valid and well-suited to the overall goal of the system, but the facial emotional expressions of the character are hard to interpret. Thus the overall design fails anyway. Or the other way around, the emotional expressions might be easily understood by the user, but the design does still not achieve its overall goal of entertaining or aiding the user.

Thus once the interpretation loop is bootstrapped and working, the second, even more challenging goal for evaluation of affective interfaces, is whether the overall usage scenarios are achieving their purpose of being e.g., engaging, fun, believable, or creating a relationship with the user, and how much of this can be attributed to the emotion modeling and expression. These two levels of evaluation will not necessarily be dividable into two different user studies or two different phases in the design process – instead they should be viewed as two levels of interpretation of what is going on when the system fails to achieve it goals.

What we are looking for, are ways of disentangling the *bad design* choices from the interesting interpretative experiences end-users have with affective systems that in many cases cannot be controlled (as they are attempting to adapt the users' emotional states and thereby changes over time) or understood in a narrow sense (as they are oftentimes portraying interesting narrative or character-based dramas).

Timing and Control As we shall discuss below, in the process of doing the studies, we found that there were some problems specific to affective interfaces that are not discussed much in the general HCI literature. These design problems concern the *timing of events* and the *level of control* handed to the end-user.

When an emotion is displayed to the user it has to come at the right point in time, and last for an appropriate length (Hendrix et al. (2000)). If an affective response from the user is the aim, then the interaction has to be carefully paced so that the user can follow it without being bored or puzzled.

As affective systems based on modelling of users' emotions are oftentimes pro-active, end-users are given less control over the interaction compared to direct-manipulation systems. The level of control and predictability needs to be balanced (Höök (1997)).

Anthropomorphism Other researcher in the field also discuss the issue of *anthropomorphism*, which can be seen as a positive or negative effect of affective interaction – in particular when realized through characters in the interface. Synthetic characters tend to raise expectations of anthropomorphism of the system (Reeves & Nass (1996)). Such anthropomorphic effects seem to have many dimensions. On the one hand the user may expect the system to be intelligent and cognitively potent. Brennan and Ohaeri (1994) showed that users talked more to the anthropomorphic interface. King and Ohya (1995) showed that users attributed more intelligence to anthropomorphic interfaces. Koda and Maes (1996) showed that realistic faces are liked and rated as more intelligent than abstract faces.

Opponents of synthetic characters argue that raised anthropomorphic expectations may lead to frustration in the user when the system cannot meet the expectations (Shneiderman (1997)). For instance, the presence of a talking face might influence the user to expect the system to possess natural language and dialogue competence, which no system of today can live up to. The general conclusion is that the more 'natural' the interface, the higher expectations on intelligence in the system. The problem arises when there is a mismatch between the users expectations and the systems' ability and this causes the user to fall out of their 'suspension of disbelief'.

Using Existing HCI Methods It should be noted that our contribution here is not an entirely new method for interactive design of affective interaction systems. We are simply picking up the methods existing within the field of HCI and attempt to see how they can be applied to this area. Thus, in the first study of the SenToy device, we used the well-known 'Wizard of Oz'-method. In the Agneta and Frida study, we used questionnaires and open-ended interviews. The Influencing Machine studies were typical laboratory-based video-recorded encounters with demo versions of the system.

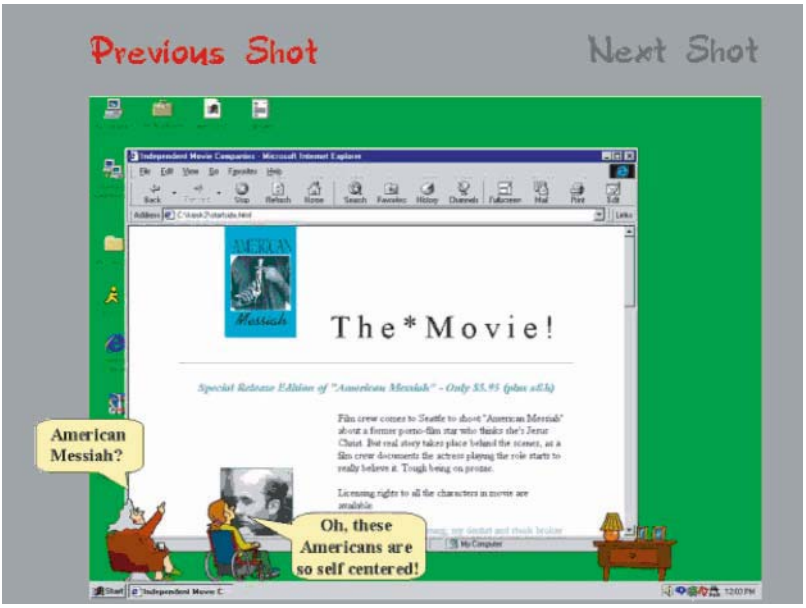
Summary of Proposed Method In summary, the method we propose is to:

- bring in end-users several times during the design work;
- apply methods that allow for a rich interpretation of users' experiences of interacting with the system;
- separate the understanding of emotional input/output from the overall experience and success of the design;
- not average over some non-existent 'normal' user, but to bring in a richer understanding of the users' background into the interpretation of what is going on between user and system;
- put some extra attention to issues of timing, control of interaction and effects of anthropomorphism (positive and negative) when observing user behavior, as well as any gaps that cause end-users to fall out of their 'suspension of disbelief'.

3. Studies of Three Affective Interfaces

The studies of the three different affective interaction systems, each illustrate a step in designing and to some extent evaluating the overall effects of affective interaction:

- The study of Agneta & Frida shows the importance of interpretation of the subjective experiences of affective systems and the risk of taking too simplistic measurements. It also shows the need to further study control and timing, and to be more open to how users' background and personality matters.
- The two studies of the Influencing Machine show the importance of first making sure that the affective output from the system is understood by users, before checking if the overall interaction idea is succeeding, thus showing the value of the two-tiered evaluation cycle. It also points to problems with control and timing, and the need for interpretative methods of analysing user study results.
- The studies of SenToy also illustrate how a study in an early stage of the design cycle can help bootstrap the design of affective input (performed through gestures with a toy) and how the second


Previous Shot
Next Shot
Independent Movie Companies - Microsoft Internet Explorer
AMERICAN Messiah
American Messiah?
Oh, these Americans are so self centered!
My Computer
Start
Independent Movie C

The*Movie!

Special Release Edition of "American Messiah" - Only $5.95 (plus s&h)

Film crew comes to Seattle to shoot "American Messiah" about a former porno-film star who thinks she's Jesus Christ. But real story takes place behind the scenes, as a film crew documents the actress playing the role starts to really believe it. Tough being on prozac.

Licensing rights to all the characters in movie are available.

Figure 5.1. Agneta and Frida reacting to the site of a film production company.

> level of evaluation can address the overall purpose of the affective interaction system. It also shows the need to differentiate between 'natural behaviors' and how users really will interact with designed artifacts.

Since we did the Agneta & Frida study before the other three studies, we shall start by describing it and the inspiration we gained on study methods. In many ways, the flaws of this study are the basis and inspiration to how we set up the studies that followed.

4. Agneta & Frida

Our first study was of the Agneta & Frida system (Höök et al. (2000)). The two animated female characters – mother and daughter – sit on the users' desktop, watching the user's browser more or less like watching television, see Figure 5.1. They make humorous and sometimes nasty comments of the web pages, the user actions, and sometimes just randomly talk to one-another.

Initial testing helped us find the right timing for the jokes – a crucial aspect of humor is to deliver it at the right moment. The early version was too slow in delivering the jokes and in particular the punch line.

Users would move on to other web pages before the joke was finished, and sometimes this meant that the joke became unintelligible.

In the following study of Agneta & Frida, we measured how many times users smiled or laughed, the amount of time they spent with the system, their mood before and after using the system, and their responses to questionnaire questions after their session with Agneta & Frida. 20 subjects tested the system with Agneta & Frida, and for comparison we also had 20 subjects who surfed the same set of web pages but without the company of Agneta & Frida.

4.1 Non-correlation of Measurements

Interestingly, none of the measurements correlated. Subject 16, for instance, smiled as often as 7.5 times per 10 minutes, spent 36 minutes (9 minutes above average) with the system, which would indicate that he had a good time. However, his post-usage view on Agneta & Frida's commentaries was only 3 on the 7-grade scale (where 7 was the highest grade). On the other hand, subject 1, who smiled the least, only 1.2 smiles per 10 minute, and only spent 16.5 minutes with the system, really liked Agneta & Frida – giving them grade 6 on the 7-grade scale. This might be because the measurements were bad and fuzzy, or because people are generally known to behave in a socially desirable way, i.e. according to what they believe the experimenter desires. But another way of explaining the non-correlation is to assume that the variables simply measure different things. We believe that although all of them try to capture the overall experience of the system, they may, in fact, measure different aspects of this experience.

For example, facial expressions of the subjects (how often they smiled or frowned) may provide indications of the immediate, un-reflected appreciation of the system. Facial expression will perhaps show the instantaneous reactions to the jokes, but not the retrospect overall appreciation of the whole experience of surfing together with Agneta & Frida. The post-usage replies, on the other hand, might reflect subject's 'afterthoughts' about the system, which may be influenced by moral and ethical preferences – the more official views of what humor and entertainment should or should not be according to a person's value system. This was in part confirmed by results such as the correlation we found between how much subjects were disturbed by Agneta & Frida and their web and computer experience. Users who had a lot of web experience were also more disturbed by Agneta & Frida ($r=.54$, $p <.05$), the same for computer experience ($r=.60$, $p <.05$). Computer experienced users may have a task-oriented and quite strict model of how to interact with computer interfaces and web browsing. Since Agneta & Frida blatantly break with this 'tradition', experienced users are more disturbed than users who do not have such strong expectations or 'preconceptions'. Especially subjects who are used to having complete control over the computer – from the insides of the operating system and out – may find it hard to accept characters in the interface and processes that run outside their control. In fact, before, after, or even during the session, some subjects said that they in general disliked interface characters for many of these reasons.

The mood measurement – which lands somewhere in-between the instantaneous reactions during use and the post-usage replies – will again measure something else than immediate reaction or the post-usage reflective evaluation. Since it showed that the Agneta & Frida subjects were in a better mood after the study compared to the subjects who surfed without Agneta & Frida, it provides us with some evidence that the system positively influences users' experience of the system on an emotional level. But being in a better mood does not necessarily mean that we appreciate every aspect of it. Our views on humor are reflections of our personality and who we want to be in the eyes of others. Sometimes Agneta & Frida make strongly ironic and sarcastic remarks about the computer and web culture, as for example:

> **Frida:** They say that computers save so much time. But sometimes I wonder... At work I often feel like I'm spending 90% of the time getting the damned thing to work, and about 10% of the time actually accomplishing things with it....
>
> **Agneta:** I don't really know... I'm not that experienced...
>
> **Frida:** Maybe we should buy a home computer...? Just for the fun of it...
>
> **Agneta:** Naa, I'd prefer a television set instead... there are more stories on TV....

Some jokes are concerned with the male dominance of the IT-world:

> **Frida:** Stupid! Nothing works! Who would ever publish a page like this?
>
> **Agneta:** A man?

Users might approve or disapprove of this type of humor or the views of Agneta & Frida. In order to determine and predict such processes, we would need a thorough investigation of subjects' attitudes towards humor, irony, and fictional characters in general, and attitudes towards these phenomena in interfaces in particular.

What aspect of experience is most important – and thus determining the appropriate method of measurement – is of course dependent on the design goals. If we aim to entertain for a onetime usage situation, then maybe it is more important that subjects smile a lot; if we want subjects to return to the system, then their post-usage evaluation should be emphasized. The fact that many users were disturbed by Agneta & Frida – but still enjoyed their company – indicated that we failed to create a feeling of flow or relaxed relationship to the space. If that had been our design goal, then other design solutions need to be sought.

Our results point at the difficulty of gathering facial expressions and using those as a means to measure subjects' affective reactions towards computer systems. Users' physical reactions of interactions with systems are not necessarily good predictors of users' inner mental states. In order to pinpoint finer distinctions in the emotional reactions, we have to consider the users interpretation, understanding, attitudes, and expectations of computer culture. The experience of jokes and irony, for instance, will be determined by personal expectations, but also by social and cultural context. As argued above, our views on humor are reflections of our personality and who we want to be in the eyes of others.

4.2 Narrative Experience

The most important design goal for the Agneta & Frida system was an idea that end-users would tie together the web surfing experience into a coherent whole: a story that would entail both the web page content and the jokes of Agneta & Frida nicely intertwined and thereby helpful to the end-user as a means of remembering the information space in a narrative form rather than as a spatially organized information space.

Apart from the measurements above, we did two kinds of analysis of the open-ended interviews performed after they had used Agneta & Frida. We asked the subjects to describe what had happened while using the system. Inspired by Maglio and Matlock (1999) and Lakoff and Johnson (1999), we performed a *metaphor analysis* of the interviews. From Maglio and Matlock's study we knew that web browsing is often perceived as a spatial activity: the user is viewed as an agent moving through the space of sites and web pages. Maglio and Matlock found this by examining the metaphors used when subjects described their surfing through web pages: 'I browse/surf the web'; 'I go to pages'; 'I enter/leave pages'; 'pages contain information'; 'the web is an information space in which I look for things'.

We decided to follow the method used by Maglio and Matlock, focusing on narrative versus spatial verbs and adverbs in the interviews that followed after out subjects had explored the system. The metaphor analysis revealed that the group of subjects who had encountered Agneta & Frida tended to talk about their experience in terms of narrative verbs and adverbs (68% narrative), while the group of subjects who only surfed the web pages without Agneta & Frida, used more spatial verbs and adverbs (only 45% narrative). The difference between the conditions was statistically significant (Mann-Whitney: $p>0.95$).[3] This seemed to indicate that users actually merged the narrative and the spatial structure into one experience. A qualitative analysis of the interviews, however, sketched a somewhat more complex picture. Subjects in our study did not gracefully merge Agneta & Frida and the web content into one narrative whole. Sometimes they enjoyed the contents of the web pages, sometimes they were amused by the comments by Agneta and Frida, and at some points web browsing and interaction was integrated into the story of the two characters, but mostly subjects divided these experiences into two separated experiences of what was going on.

Finally, we measured *disturbance* and *recall*. If the user was able to integrate the narrative of Agneta & Frida with the web content, we hypothesized, that subjects would be less disturbed by the two characters, than a case in which the Agneta & Frida story ran 'in parallel' to the web content. In the latter case, the comments and activities of the characters would be experienced as intrusive. As for recall, we assumed that the emotional reactions caused by the remarks from Agneta & Frida – e.g., laughs, frustration, moral judgment and agreeableness – would enhance the recall of the information remarked upon. We assumed that Agneta and Frida would encourage the user to construct a narrative context and associative links between information in the site, which would improve memory. Thus, we expected the Agneta and Frida subjects to perform better on a post-usage recall test, than would subjects without Agneta & Frida.

There was no difference between the two groups in terms of how much they remembered of the web pages. Out of the 38 randomly selected test pages, the Agneta & Frida group remembered 88% of the pages they had seen, while the group who surfed without Agneta & Frida remembered 89%. Subjects were able to accurately recall the comments Agneta & Frida had made at particular pages. It seems like Agneta & Frida failed to create the context needed to better tie the different sites in the space together into one coherent narrative experience.

4.3 Implications for Design Method

While these results basically only tells us that the design was bad in terms of achieving this particular goal (even if Agneta & Frida were indeed successful in many other ways), the results also tell us something really important about the need for open interviews and deep interpre-

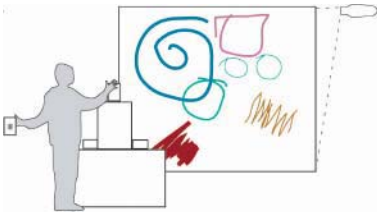


Figure 5.2. Setup of the Influencing Machine.

tation of what is really going on between subject and system. Simplistic measurements of time spent, bodily reactions, or questionnaires will only provide a limited understanding of what users really feel and think about complex, interactive systems such as Agneta & Frida. In fact, if we would have decided to only tell the story of those measurements, then Agneta & Frida would have looked like a very successful system.

Second, we also learnt how crucial the background and subjective perspectives of end-users were in how they reacted to Agneta & Frida. For any evaluation of a desk-top program, end-users values, humor or personality would not be considered crucial to how they react to the system. In this case, those aspects became key.

The level of control given to the users was also crucial to some of the subjects. The computer-experienced users did not like the lack of control that they experienced when Agneta & Frida interrupted their interaction and acted independently.

5. The Influencing Machine

We took many of the experiences from the Agneta & Frida study with us when we studied the Influencing Machine designed by Phoebe Sengers and colleagues (Sengers et al. (2002), Höök et al. (2003)).

The Influencing Machine explores the tension between machines and affective beings in affective computing; how people will relate to a machine whose emotions they can influence, but whose behavior they cannot control. In some ways it can be seen as a provocative piece of interactive art exploring some of the Artificial Intelligence (AI) dreams with a more critical, cultural perspective.

The Influencing Machine is supposed to work as follows. Two people enter a small room. Child-like scribbling appears across a wall: jagged lines, circles, spirals, and other shapes build up, overlap, fade away (see Figures 5.2 and 5.4). Scattered throughout the room are postcards with art prints or color fields; on a table stands a wooden mailbox (see

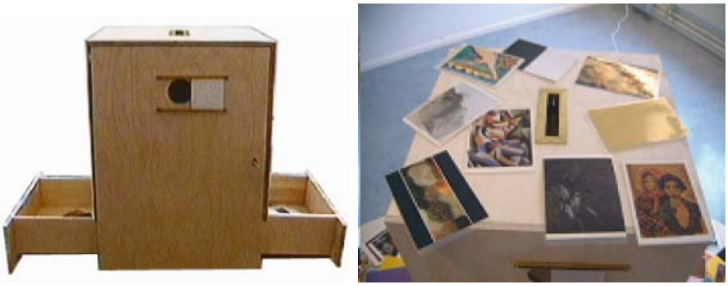


Figure 5.3. The constructed Mailbox.

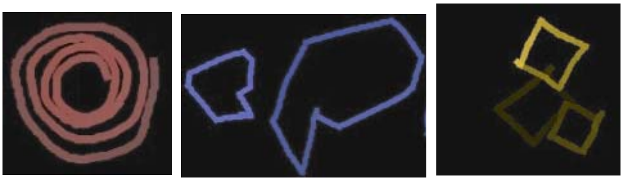


Figure 5.4. Examples of generated scribblings.

Figure 5.3). One person picks up a card and tentatively puts it in the box. Unusual and musical sounds begin to play. Drawings change speed, color, pressure, form. The people begin sorting through cards, dropping them in the box and seeing how the graphics and sound change. They play, experiment, and discuss: "How is this reacting to us?" "How do you think this works?"

Technically, the system works by using the input postcards marked with machine-readable bar codes to influence an internal emotional model. These internal emotions trigger sounds and the selection of drawing behaviors and their dynamic parameters: speed, color, size, pressure, etc. When the machine receives input, system drawings tend to become gradually more complex; when it has not received input for several minutes, it restarts. While this technical description is precise and clean, the emotional interpretation of the graphical output and postcards by users is complex, incompletely specifiable, open-ended, and strongly culturally influenced.

5.1 Study Method

The *co-discovery method* (Dumas and Redish (1993)), where users are brought in two by two, was used with some slight modifications. We brought in users in different group sizes. Also, we were not interested only in the talk-aloud effect, but also in group dynamics around the art piece. Facial expressions and discussions among the subjects are much more interesting to study with a group of users as opposed to single users in front of a screen. Second, we asked more questions about the subjects' background and attitudes than in the Agneta & Frida study. Thirdly, we kept the interviews after their session much more open-ended to allow for them to express various views and ideas, rather than a simple "Yes, I like it" or "No I don't" in a questionnaire.

Agneta & Frida and the Influencing Machine are quite different systems. The Influencing Machine grew out of the affective computing field, but takes on a different stance. Affective computation generally focuses on the informatics of affect: structuring, formalizing, and representing emotion as informational units. Through the Influencing Machine Sengers and colleagues proposed instead an enigmatics of affect, a critical technical practice that respects the rich and undefinable complexities of human affective experience. The Influencing Machine bridges the subjective experience of the user and the necessary objective rationality of the underlying code. It functions as a cultural probe, reflecting and challenging users to reflect on the cultural meaning of affective computation. In doing so, it might not aim to please, as Agneta & Frida did, but instead to spur reflections and discussions.

But what exactly were we going to check once we brought the Influencing Machine and users into the lab?

The purpose of the Influencing Machine is to create a cultural provocation, challenging our views of what a machine can be, in particular whether it was capable of being emotional – but how would we check what the machine in fact was able to provoke? What if users did not get the idea at all, or if they only got frustrated and dismissed it entirely? A provocation entails an experience that is not necessarily easy or pleasant for users, so we may have the goal of developing painful or difficult situations. This is something standard usability strategies will try to avoid.

We had to disentangle frustration that came from *bad design* choices from frustration that came from actually encountering a machine that cannot be controlled – only influenced. The design of the Influencing Machine is balancing on a thin line between being predictable and controllable and thereby boring and not achieving its purpose, and being unpredictable and uncontrollable and thereby alienating its users, making them feel stupid and out of control entirely.

5.2 The First Influencing Machine Study

Our first study of the machine was done at a very early stage in the development cycle. The Influencing Machine did not have any sound system at this point. The evaluation was explorative in nature, as our main goal was to feedback into the design process.

Users were brought in small groups (six groups with in total 12 subjects) into a room with the Influencing Machine. Users were told that the installation had something to do with emotions, and were then allowed to play with the system as long as they liked. On average, they spent about 20 minutes in the room.

Generally speaking, users were first curious, then became frustrated. Often this frustration stemmed from not being able to control the machine. They had a great deal of trouble figuring out the relationship between postcards and drawings. For some users this became a barrier that stopped their interest in the machine. Some users found the Influencing Machine drawings too simple and drawn too slow. The mailbox itself was liked. Unfortunately, the bar code reader in the mailbox made a beep whenever a postcard was inserted. This led subjects to think of the mailbox as a machine rather than a form of communication with a semi-living being.

A complication was the frustration that users often developed with lack of control. Many users got irritated and frustrated when they could not figure it out. Certainly this is an affective reaction, but not one intended, unless leading to the kinds of discussions sought by the designer/artist. These thoughts and observations led to a number of system design changes performed by Sengers and colleagues.

Users were confused about the emotional meaning of the imagery. The addition of the sound system helps to clarify the agent's interpretation of input cards and its emotional state. Moreover, an internal emotional display was developed showing the level of each of the internal emotions. Although the designers of the Influencing Machine were reluctant to show these internals, by offering the user an opportunity to understand how the agent is designed to feel, users can and do engage in critical reflection on whether they believe that the drawings actually express the stated internal emotion state. This display can be set in a state were it will fade away over time, supporting users through their initial exploration without constraining further interaction.

Users were also confused about the nature of influencing versus controlling the system. With the above improvements to emotional expression, including direct sound feedback instead of mechanical Mailbox beeping for changes in emotion, users would hopefully have a better understanding of how they affect the system. At the same time, this concept is subtle and runs counter to users' everyday experiences with computers; it may simply be in its nature that it is hard for users to understand.

Finally, users were sometimes bored by the drawings themselves. Speeding up the drawings, reducing the persistence of behaviors so that new forms appear more quickly, and adding some more complex drawings will probably raise user interest. Also, transitions between drawings need to be handled more gracefully. In the old version, the system draws for a while and then clears the screen and starts over. The graphics was re-implemented to remove these rough breaks by layering over one another and gradually fading away.

In general, the first study achieved the first level of feedback to the design envisioned by our two-tiered design method discussed above. It made clear what aspects of the affective input means and the affective output from the system were understandable/failing to the users preventing them from going from a 'basic' level of understanding the input – output relation, to actually starting to reflect on the overall purpose of the machine.

5.3 The Second Influencing Machine Study

The second study was performed in a similar fashion to the first study, but on an improved and altered Influencing Machine. In this new version, the timing was faster, the scribblings more complex and interesting, and an explanatory 'emotion bar' was added to the top of the scribblings showing the emotional state of the machine.

The results from this second study showed that the design changes did indeed achieve the desired result; users were more positive, less confused, and more of them did understand the point and were willing to discuss the intended provocation than in the first study. The replies to the interview questions and the interactions the groups did with the machine indicated that the group who had the emotional display on did more easily grasp that the machine expressed emotions and could be influenced.

Subjects were more inclined to form theories of what was going on inside the Influencing Machine and we got more positive comments about the drawings and the overall experience. The subjects from the second study also used the Influencing machine twice longer in average than the subjects from the first study. But there were still those subjects who experienced frustration and who were less inclined to 'get the point'.

5.3.1 Video Analysis The analysis of users' experiences of the Influencing Machine was done through carefully transcribing everything that the subjects did with the machine as well as their dialogue with one-another and not to avoid interpretation of what was going on and how subjects' personality interfered with their interaction. We will discuss the case of group 6.

#6 Two Teachers and a Husband The three subjects were 61 (female), 65 (male) and 42 (female) years old. The two women were teachers, and one of them, was married to the man.

The two women did not look very carefully at the cards that they put in the machine. Nor did they analyze what was happening on the screen. The machine restarted after 3 minutes. Both women kept on entering cards very quickly. The man was quiet, kept to the background, and only gave away something of his theories after about 8 minutes. In general, one woman, his wife, was quite dominating and the man had a hard time convincing her that his theories could be proven. The two women realized that the machine kept on drawing even when they did not put any cards inside the machine, and used this as an argument that the man's theories could be dismissed.

The man did not give up, but discussed the emotional display and said that one has to put a card inside the machine in order to make the values in the emotional display fluctuate. He got some positive feedback on his theory from the machine, and albeit reluctantly, he got the two women to take part in some more theory forming. Unfortunately, the machine did not react to the postcard that the dominant woman inserted, at least not visibly. The man got more visible reactions to his postcards, which in turn made him think that the machine only reacted on him. He suspiciously turned around, staring at the video camera, wondering whether this was in fact where the 'control' was placed.

During this, the dominant woman made an interesting comment: she pointed at the computer under the table with the table cloth, and asked the man whether this computer was in fact connected to the machine. She meant that if it was, then the Influencing Machine was just a computer – not a machine in its own right. It seemed that to her a computer cannot be what she perceives that the Influencing Machine is (according to the man's theories). If it is a computer, it must be predictable, not influenced by them.

They stopped putting in cards for a while which caused the drawings to change color until they were white and the machine restarted. They put a few cards inside the machine and then they waited for it to restart again, just to see if the drawing would change color to white again before the machine restarted. Again, the man argued that the cards they put in the machine seemed to be influencing it, but the other two argued that the card is not important and that the machine just went around in a cycle: "placed on 'repeat"'.

They waited for the machine to restart a third time, to check if the machine would start drawing even if they did not insert any cards, and they found that it did. They discussed whether the machine would restart if they stopped inserting cards or if it restarts after a certain time interval. They speculated about whether the emotions were connected with certain colors in the drawings. Finally, the dominant woman concluded that it was entirely random, while the man kept on insisting that there were certain relationships to his actions.

This summarised transcript shows how theories were formed and discussed, and how the Influencing Machine was even capable of spurring the kind of discussion of what a computer can/cannot be that the designer/artist sought.

In total, seven of the nine groups invented different theories that they tested during their session with the machine. They tried to make the machine respond in a particular way by putting a certain card or a specific category of cards inside the machine; for example, they tried to use only dark-colored cards in order to see the response from the machine. The groups that tested several different theories during the session seamed to have more fun during the session than the other groups, but after a while most of them got frustrated when the response from the machine was not what they expected.

5.3.2 Timing and Control In the Influencing Machine, the *timing* of emotion change and development, drawings, and system's reactions to inserted postcards is key. The interaction cycle must be slow enough for users to recognize the emotions, but fast enough to attract and keep the users' interest. The intent is not for the user to control the machine, but also not to make users too frustrated when they cannot control it at all. The second study showed that the design of the machine was closer to a reasonable balance point.

5.4 Implications for Design Method

The two studies of the Influencing Machine showed the need for in-depth interpretation and analysis of users' behavior. The study is an explicit attempt *not* to avoid the messiness of having several users together in the lab, interpreting their behaviors based on some subjective understanding of their personality and attitudes. Through such a study, we could give the designer of the Influencing Machine a *grounded feeling* for what works.

The study also showed the usefulness of first making certain that the affective input – output behavior could be understood, before studying the overall design against its purpose.

Problems that reappeared in this study had to do with perceived level of control – a natural consequence of provoking users preconceptions of machines as stupid, rational, and predictable – and timing of the affective behaviours. These two factors are not unrelated. After speeding up the response from the machine, users felt that that they could understand and control the machine to a larger extent than in the first study.

Laboratory evaluations helped us uncover problems in interaction design related to questions like: "Is this interaction cycle right? How is the timing? Do users understand the affective expressions?" In the case of the Influencing Machine this meant reaching the balance point between control and complete randomness (in the eyes of the users), finding good timing so that users are captivated (and not bored), finding the right level of interesting drawings, and getting better sound.

Finally, let us point out that evaluation of this kind can give answer to the question "Is it good interaction?", but not to the one "Is it good art?" If our question is "Is it good *interactive* art?," we may need to more fully integrate the perspectives of art and HCI. We suggest this may be done by a 'system critic,' who analogous to a literary, movie, or art critic is specialized in understanding the social, cultural, and intellectual context of the system and who simultaneously can evaluate the system using variations on standard HCI techniques.

6. SenToy and FantasyA

Finally, the last system we have designed and studied was an affective input device – the SenToy – and a game named FantasyA (Andersson et al. (2002), Paiva et al. (2003), Höök et al. (2003)). SenToy is a doll with sensors that allows users to (partly) control their avatars in an adventure game. SenToy allows players[4] to influence the emotions of a synthetic character placed in FantasyA, a 3D virtual game. By expressing gestures associated with anger, fear, surprise, sadness and joy through SenToy, players influence the emotions of the character they control in the game. Players' characters will be drawn into duels where the expressed emotion determines which spell is cast at their opponents, the players' character will trade (using emotion expressions) with other characters to win magic stones, and so on.

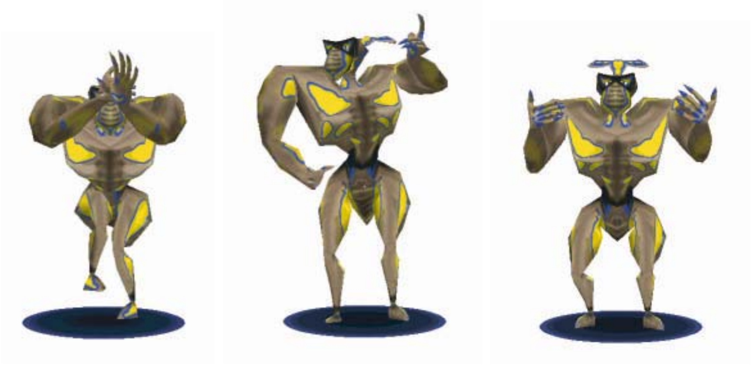


Figure 5.5. Fear and two versions of Gloat as expressed by one of the avatars (stills of animated behaviour).

The aim of SenToy is to 'pull the player into the game' through the use of a physical, touchable affective interface. With sensors in its limbs, sensitive to movement and acceleration, SenToy is designed to capture certain manipulations patterns from players, which in turn are associated with particular emotional expression.

The affective output in the system is shown through how the avatar that the player controls behaves, see Figure 5.5. This in turn also determines what the character will do next. Emotions as expressed through SenToy, controlling the avatars emotional state and subsequent actions is therefore the only way that the player can play the game.

6.1 Wizard of Oz

When designing SenToy it was hypothesized that players would manipulate the toy to express emotions by using a particular set of gestures. Those gestures were drawn from literature on how we express emotions through bodily movements and from emotion theories (Darwin (1872/1998), Davies (2001)). To evaluate this idea we performed a Wizard of Oz study (Andersson et al. 2002). Wizard of Oz studies have previously been used for natural language interface (Dahlbäck et al. (1993)) and intelligent agent design (Maulsby et al. (1993)) and we showed that it can effectively be used also in the domain of affective input design.

In a Wizard of Oz study, users are made to believe that they are interacting with a system, while in reality they are interacting with a human Wizard, sitting behind the screen pretending to be the system. This study was performed with dolls that did not have any sensors at all, but where the Wizard interpreted users' actions with the doll and made the avatar express the corresponding emotion. Since subjects divided their visual attention between the doll and the screen with their avatar, subjects sometimes missed the actual performance of an emotion of the avatar's face or body as they were focusing on the doll and moving the doll. The Wizard adjusted to this problem, delaying until the subject had finished their movement with the doll, or sometimes, even making the avatar perform the action twice.

The study showed that there are movements with the doll that most users will easily pick up to express emotions, but that these are not necessarily linked to any 'natural behavior'. First, users will not behave in the same way when expressing emotions through a doll rather than through their own bodily behaviors. There are numerous reasons for this, among them the cultural notions for how dolls and cartoon characters behave when expressing emotions. Secondly, we needed to put users in a loop where they are given feedback from the system through how the avatar reacts. Users will learn how to create the right behavior through watching the face of the avatar when they perform actions on the SenToy. Thus there is room for 'unnatural' learnt behaviors. In addition, imitation between avatar animation and end-users' movements with the doll, will probably take place (and did in fact happen during the last study).

The WoZ study also revealed some other aspects of the design of the doll and its interaction through the sensor technology, such as the preferred distance between user and screen, movements of limbs that

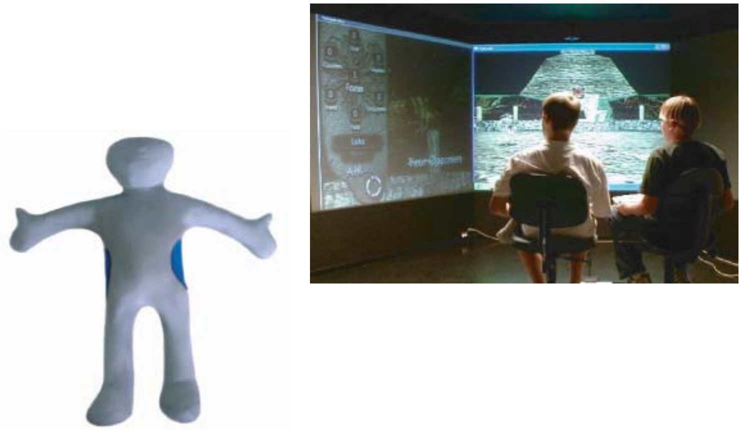


Figure 5.6. SenToy to the left and boys playing the FantasyA game through SenToy to the right.

will most likely occur, desired softness and size of the doll, and which facial expression it should have (neutral).

Based on the results from this study, the doll in Figure 5.6 was designed and implemented. The movements for each emotion are described in Table 5.1.

Table 5.1. Mapping of emotions to recognized expressions.

Emotion	**Expression**
Happy	Jumping/Dancing up and down
Sad	Lean the body forward at least 45 degrees
Gloat	Point right arm forward and jump up and down
Anger	Shake doll forwards and backwards or side to side
Fear	Hand(s) in front of eyes
Surprise	Jump back rapidly, and tilt backwards at end

6.2 Second Study of SenToy Used in FantasyA

In the second study of SenToy, we were able to use a functioning prototype of the toy based on the movements collected from the WoZ study and an early version of the adventure game named FantasyA. Users (players) were brought in as pairs and were encouraged to play together. In general, the conclusions were that SenToy was a great success, but that some of the emotions did not necessarily make sense in the context of the game. The game itself was also quite complex and only a few of the players did understand what was going on.

Subjects found it fairly easy to express most emotions, with the exception of the emotion Surprise. Surprise was also only rarely used during the game. The most used emotions were Gloat and Happy, on second place came Sad and Angry, on third place Fear, and finally, Surprise.

During the game most emotional expressions were very physical and encouraged players to act out the emotion. The exception from this rule was Sad where subjects sat very still, bending the doll over waiting to see the result on the screen. This is not necessarily a bad design choice since sadness is characterized by an inwards posture among people, thus encouraged by the design of the movement.

Some users, especially the kids, were really keen on having the doll and would pull it from the other player or interfere and try to help the other player in expressing some particular emotion. In the interviews, two kids commented that they would have liked to have a doll each and be able to play against each other.

In the comments field of the questionnaire, one player wrote:

> A few days after having played, I still like the doll very much. I really appreciated his direct contact to give commands, even if in that case, the commands were not that obvious and their result a bit fuzzy. (adult player)

One of the kids remarked that he would probably like to use the SenToy for a whole month before getting bored. Considering that he was 12 years old, this is a very good result.

After the game about 80% seemed to like the doll. The kids were in general more enthusiastic about the doll than the adults. In the interviews about the SenToy some players felt that they became one with SenToy, but others felt that a button-based interface would have made them feel more directly in control of it. In general, the impression given was that they could identify with the doll most of the time and act through it, but that the avatar was reacting in strange ways sometimes, thus they did not feel that they through the doll became the avatar.

Players also seemed to have an intellectual rather than emotional relationship to the emotions of their own avatar and to the emotions expressed through SenToy. They would "instrumentalize" the emotion to be one of the commands in the game, such as "cast blast" or "cast shield". They would be playing the strategic, intellectual game rather than being influencing on a basic instinctive emotional level. This was due to several different design decisions – some of which might be easily changed if the aim is to make the player more emotionally affected by the game.

On another level, players do get more and more involved with the game – especially when they win a few duels – but to the experimental leaders this seemed to be more in terms of "duel emotions" than the six emotions that can be expressed through the doll.

The FantasyA game is currently being redesigned by Paiva and colleagues to better cater for an emotional involvement between user, SenToy and their personification in the game as their avatar. The narrative structure connecting the game turns with the emotional states of their avatar will be the key to further developing the game, together with these study results.

6.3 Implications for Design Method

The design and user studies of SenToy and FantasyA show how user studies can be very relevant to do even before a system has been implemented or fully designed. The Wizard of Oz study saved a lot of energy in the project through pointing out the flaws in the theory of how people would move the doll to express different emotions. The design of SenToy, similar to the design of the mailbox in the Influencing Machine, also show how these affective interactive systems are indeed designed artifacts with their own interaction problems that cannot be solved simply through creating an even better theory of human emotions and emotion expression. Arriving at a good affective game or an interesting affective interactive art piece, is a process where the user studies can help to debug the particular interaction functions.

While not used as much in the studies of SenToy and FantasyA, subjective evaluation and interpretation of what where experiencing when using the system were crucial. It is through such an interpretative analysis that we could see that users did not identify directly with the emotions they were expressing in the game, such as sadness or surprise, but that they instead were reacting with a different set of emotions much more related to their game play experience. We believe that a careful analysis and redesign of the relationship between emotion and the next game turn could create a system where the two are more in harmony and players will start to experience the emotions they are expressing through the SenToy.

The two studies of SenToy and FantasyA again show the importance of first getting the affective input – output relationship right before attempting to evaluate, in this case, the success in terms of how well the affective game captures users' interest and achieves affective involvement. Since the design of this system is not yet finished, yet another study would probably be a good last step in the design cycle.

7. Discussion

The studies of the three different systems show the importance of bootstrapping affective interaction and making sure that the affective expressions or affective input opportunities are understood before the overall system can be evaluated. The studies also reveal some important issues to be dealt with once this bootstrapping has been done and the system is evaluated against its overall purpose. In particular, we find that the field of affective computing often make simplistic statements where it is claimed that e g users will more easily bond with an affective system, become more efficient if not stressed or disturbed at the right moment. The Agneta & Frida study and the Influencing Machine study show how complex the reactions are to these interfaces and how much depends on the users' background, age, attitudes and interest – to some extent this is different from normal usability issues.

Some general conclusions about design difficulties of affective interfaces can be drawn. First of all, all the studies confirm that issues of timing are crucial. Agneta & Frida's jokes have to arrive at the right moment, the Influencing Machine has to be influenced at the right level and draw its drawings fast enough in order for the interaction to work, and finally, the avatar reactions to SenToy has to be delayed or prolonged enough for the user to both handle the doll and watch the avatar on the screen in order to understand what happens next in the game.

For SenToy, many lessons were learnt before the costly process of creating a doll with sensors was started. All studies, but in particular the SenToy study, definitely show that it will be a mistake to only aim for "naturalness" in the affective expressions. From the theory of human expression, a set of movements were extracted, but in the two studies, these movements were not the ones that best fitted with the particular game situation and how users did really behave with the doll. Most interactive agents and affective interfaces are interesting in that they are different from how we behave in human-human relationships, but still similar enough for us to recognize them and have fun with them. This concurs nicely with theories such as those presented by Suchman (1987, 1997) or by Dourish (2002). Dourish argues that rather than embedding fixed notions of meaning within technologies, we should allow users to create and communicate meaning through their interaction with the system and with each other through the system, since this is how artefacts are given their meaning in human culture.

We need to do more of these open-ended explorative studies, early on in the design process, before we can start doing the studies that really matter: namely those that show that affect in interaction does indeed contribute something different from other kinds of design. In this process, we need to more openly discuss which measurements will indeed be related to the overall goals of the entire system. The non-correlation between measurements in Agneta & Frida shows how difficult it is to separate an understanding of what kind of experience we want to evoke from users' attitudes and values. It also shows that we need to be clear of what kind of experience it is that we want to give the user: a short-term fun thing, a post-usage positive attitude, a provocation that continues even after using the system as for the Influencing Machine, or what?

While we have not presented a complete framework for how to bootstrap design and evaluate affective applications, we believe that our studies could be the inspiration to taking some more steps in this direction. In particular, we hope to encourage taking users into the loop when designing the interaction cycle with respect to timing, narrative context, understanding of affective input and output, and being more open to the effects of users' attitudes and cultural values.

Acknowledgments

Thanks to Gerd Andersson, Adrian Bullock, Pia Mårtensson, Per Persson, Ana Paiva and Phoebe Sengers who worked with me on the studies discussed here. Thanks to Petra Fagerberg and Anna Ståhl for reading and commenting this text. Thanks also to the anonymous subjects who participated in the studies. Most of the work was done within the EC-funded SAFIRA (Supporting Affective Interactions for Real-time Applications) project while the author was employed at SICS.

Notes

1. It should be pointed out that though the author was involved in all the user studies described in here, the designs and studies of the systems were performed by teams of researchers.

2. Believability refers to how well those characters are able to appear as living, coherent characters that users are willing to interact with.

3. An evaluation of a web-interface from 1994 done by users 2004 would tell us that it looks boring, old and unusable, has all its buttons in the wrong places, does not use frames properly, while an evaluation of the same interface done back in 1994 would probably show completely different results. Computer interfaces are cultural artefacts.

4. We use the term player rather than user throughout the description of this system to emphasise that the target domain is a game.

References

Ark, W., Dryer, D., and Lu, D. (1999). The emotion mouse. In *Proceedings of HCI International 1999.* Munich, Germany.

Andersson, G., Höök, K., Mourão, D., Paiva, A., and Costa, M. (2002). Using a Wizard of Oz study to inform the design of SenToy. Exhibit at *Designing Interactive Systems, DIS'02*, ACM Press, London.

Bates, J. (1994). The Role of Emotion in Believable Agents. *Communications of the ACM, Special Issue on Agents*, 37(7): 122–125.

Brennan, S.E. and Ohaeri, J.O. (1994). Effects of Message Style on Users' Attributions toward Agents. In *Conference companion on Human factors in computing systems, CHI'94*, pp. 281–282, Boston, Massachusetts, United States.

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pp. 193–200, ACM Press, Orlando, Florida, United States.

Damásio, A. (1995). *Descartes' Error*, Avon Books, New York.

Davies, E. (2001). *Beyond Dance: Laban's Legacy of Movement Analysis*, Seven Locks Press, Santa Ana, CA, USA.

Darwin, C. (1872/1998). 3rd ed. by Paul Ekman, *The expression of emotions in man and animals.* Oxford University Press, Oxford.

Dumas, JS. and Redish, J. (1993). *A Practical Guide to Usability Testing*, Ablex, Norwood, NJ.

Fernandez, R., Scheirer, J., and Picard, R. (1999) *Expression glasses: a wearable device for facial expression recognition*, MIT Media Lab Tech. Rep. 484, Cambridge, MA.

Gaver, W., Hooker, B., and Dunne, A. (2001). *The Presence Project.*, Royal College of Art, London.

Hendrix, J., Ruttkay, Zs., ten Hagen, P., Noot, H., Lelievre, A., and de Ruiter, B. (2000). A facial repertoire for avatars, *Proceedings of the Workshop Interacting Agents*, pp. 27–46, Enschede, The Netherlands.

Hassenzahl, M., Platz, A., Burmester, M., and Lehner, K. (2000). Hedonic and Ergonomic Quality Aspects Determine a Software's Appeal. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 201–208, The Hague, The Netherlands.

Höök K. (1997). Evaluating the Utility and Usability of an Adaptive Hypermedia System. In *Proceedings of the International Conference on Intelligent User Interfaces*, pp. 179–186, Orlando, Florida.

Höök, K. (2000). Steps to take before IUIs become real. *Journal of Interacting with Computers*, 12(4):409–426, February.

Höök, K, Bullock, A., Paiva, A., Vala, M., Chaves, R., and Prada, R. (2003). FantasyA and SenToy. In *Proceedings of the conference on Human factors in computing systems*, pp. 804–805, Ft. Lauderdale, Florida, USA, ACM Press.

Höök, K., Persson, P., and Sjölinder, M. (2000). Evaluating Users' Experience of a Character-Enhanced Information Space. *Journal of AI Communications* 13(3): 195–212.

Höök, K., Sengers, P., and Andersson, G. (2003). Sense and Sensibility: Evaluation and Interactive Art. In *Proceedings of the conference on Human factors in computing systems*, pp. 241–248, ACM Press, Ft. Lauderdale, Florida, USA.

King, W.J. and Ohya, J.(1995). The representation of agents: A study of phenomena in virtual environments. In *Proc. of the 4th IEEE International Workshop on Robot and Human Communication ROMAN'95*, pp. 289–290, Tokyo, Japan.

Koda, T. and Maes, P. (1996). Agents with Faces: The Effects of Personification of Agents. In *Proceedings of Human-Computer Interaction*, pp. 239–245, London, UK.

Lakoff, G. and Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, Basic Books, New York.

Maglio, P. and Matlock, T. (1999). The Conceptual Structure of Information Space. In Munro, A. , Höök, K., and Benyon D., editors, *Social Navigation of Information Space*, Springer-Verlag, London.

Martinho, C. and Paiva, A. (1999). Pathematic Agents. In *Proceedings of the third annual conference on Autonomous Agents*, pp. 1–8, Seattle, Washington.

Maulsby, D., Greenberg, S., and Mander, R. (1993). Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the conference on Human factors in computing systems*, pp. 277–284, Amsterdam, The Netherlands.

Norman, D. (1990). *Design of everyday things, The Design of Everyday Things.* Doubleday, New York.

Ortony, A., Clore, A., and Collins, G. (1988). *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge.

Paiva, A., Andersson, G., Höök,K., Mourao, D., Costa, M., and Martinho, C. (2003). SenToy in FantasyA: Designing an Affective Sympathetic Interface to a Computer Game. In *Journal of Personal and Ubiquitous Computing*, 6(5-6): 378–389, Springer-Verlag, London Ltd.

Persson, P., Laaksolahti, J., and Lönnqvist, P. (2002). Understanding Social Intelligence. in Dautenhahn, K., Bond, A., Canamero, L. C., and Edmonds, B., editors, *Socially Intelligent Agents - creating relationships with computers and robots*, pp. 21–28, Kluwer, Dordrecht.

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

Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New media Like Real People and Places*, Cambridge University, Cambridge.

Sengers, P., Liesendahl, R., Magar, W., Seibert, C., Müller, B., Joachims, T., Geng, W., Mårtensson, P., and Höök, K. (2002). The Enigmatics of Affect. In *Proceedings of Designing Interactive Systems, DIS'02*, pp. 87–98, ACM Press, London.

Shneiderman, B. (1997). Direct Manipulation for Comprehensible, Predictable and Controllable User Interfaces. In Moore, J., Edmonds, E., and Puerta, A., editors, *Proceedings of 1997 International Conference on Intelligent User Interfaces*, pp. 33 – 39, ACM Press, Orlando, Florida.

Suchman, L.A. (1987). *Plans and Situated Actions: The problem of human-machine interaction.* Cambridge University Press, New York.

Suchman, L.A. (1997). From Interactions to Integrations. In Howard S., Hammond, J., and Lindegaard, G., editors, *Proceedings of Human-Computer Interaction INTERACT'97.* p. 3, Sidney, Australia.