Evaluating Affector: Co-Interpreting What "Works"

Phoebe Sengers, Kirsten Boehner, Simeon Warner, and Tom Jenkins
Cornell Information Science
Culturally Embedded Computing Group
Ithaca, NY 14850 USA

INTRODUCTION

Affector is an interactive installation intended to stimulate reflection on and interpretation of emotions, based on an autobiographical design method in which the same people are designers, users, and evaluators. Evaluating Affector is challenging because we cannot evaluate the system by itself, but instead need to consider the role of users, designers and evaluators in actively appropriating and altering system meaning. Here, we describe the methods we are developing to evaluate Affector, which we hope to discuss and further develop with the other participants in the Evaluating Affective Interfaces workshop.

Motivation

We have been experimenting with systems to express affect for several years (4,3,1). In evaluating these systems, it became clear to us that because of the complex and ambiguous nature of affect, human users rarely interpret the affective output of these systems the same way it is represented in the system's relatively simplistic internal emotional model. We began to ask ourselves whether internal emotional models were distracting us as designers; would it be possible or perhaps even better to build systems to express emotions without directly representing them? Could one develop a computational system that users can usefully interpret emotionally without building emotional models into the system? And in doing so, could we deal with emotion in a more ambiguous, rich, and situated way than is possible when it must be reduced to discrete categories to make it understandable to computers?

The result of these musings is Affector, an experiment in the co-interpretation of affect. A video window between the offices of two friends communicates their moods by systematically distorting the video feed according to sensor readings. Emotion is not directly represented in the system but is instead interpreted by its human users as they tune its distortions to match their intuitions of their moods. Affector's design is intended as a challenge to the push in computer-mediated communication for greater realism and accuracy in representation. The system emphasizes instead openness to interpretation and manipulation by end users.

What do we evaluate when we evaluate Affector?

Before a single line of code was written, we already struggled with the question of how such a system would be evaluated. How do you evaluate something meant to evoke reflection on and interpretation of emotion rather than to represent emotion? If our focus is on users appropriating the meaning of the system in novel ways, could we even make the claim that one system 'works' whereas another does not, or one system works 'better' than another, when in fact one could argue it is the users who are working better (or worse)?

Something as enigmatic and amorphous as affect in any case challenges existing models of evaluation. Furthermore, evaluation itself is a multi-faceted concept and some of the difficulty in evaluating Affector lies in first articulating what is being evaluated and why. Under the large umbrella of technology 'evaluation', there are three different things potentially being evaluated: the system itself, the design and design process, and the actual use or appropriation of the system by users. These three forms of evaluation ask different questions, use different methods, and are not necessarily easily compatible with one another.

In system evaluation, perhaps the most familiar type of evaluation in HCI, the guiding question is whether or not the system is successful, where success must first be clearly articulated and defined. Does the system do or enable what it was designed for? In this evaluation scenario, the system itself is primarily under review, and the goal is to improve the design, i.e. to make it work better.

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.

Finally, in evaluations of technology use and appropriation, typically used in the social sciences such as psychology, sociology, and communication, the guiding question is how people make sense of, communicate meaning about, and participate in the world. In this type of evaluation, the use of a technology for meaning-making activities is primarily under review. The goal is to describe, explore, and possibly change social, cognitive and emotional interactions.

In evaluating Affector, our challenge is to integrate these three, sometimes contradictory, kinds of evaluation. We are interested in whether Affector works and how it can be improved. But understanding how Affector works is not possible without also understanding how it is appropriated by its users, since we do not have a specific goal for what Affector should communicate at any moment, and instead

would like to see how users make sense of it. At the same time, we know that as designers we do have more general goals for the system, and we believe it is important to critically examine how our design decisions are shaping user experience and whether and to what extent we are really opening up a new space for user appropriation.

Because Affector depends strongly on user interpretation, we foreground critical evaluation and evaluation in use, yet all three types of evaluation inform each other. While the evaluation of the system fades into the background compared to usual HCI studies, it acts as a stimulus for the interplay between evaluating the design and the actual use of the system. In other words, evaluating the system is a proxy for critically examining why certain design choices were made and their effects. The guiding question is not whether it works, but how we define what works. How do we appropriate a system to make it work? And what about the system supports these definitions and appropriations?

AFFECTOR: DESIGN SKETCH

Phoebe Sengers and Simeon Warner are friends who happen to work in the same building, their offices next door to each other. One evening, Phoebe was working late when she heard someone go into Simeon's office and say, "Working so late by yourself?" Although up to then she had been unaware anyone else was there, Phoebe spontaneously shouted, "He's not by himself - I'm right next to him!" Indeed, Phoebe and Simeon's office chairs are only a meter or two apart, although the intervening wall means neither is normally aware of the other's presence. They decided on the spot that they would build a virtual window between their offices that would let them be aware of each other's emotional presence.

The central goal of Affector is to support friends in shared office spaces in maintaining an ambient sense of each other's moods. The system should require little active intervention; it should communicate a background sense of mood autonomously, rather than being told by the office residents what they should communicate. The system should not directly model user emotions, understood as discrete and well-defined units, but rather give a continuous, rich, and potentially ambiguous background sense of emotion. Disambiguating system output is in the province of the systems' users, drawing on the friends' existing rich understanding of one another based on their day-to-day interaction.

There are many existing systems which function effectively to support ambient awareness in office contexts. Affector differs from these in two major ways. First, most systems for awareness in computer-supported cooperative work (CSCW) are intended to improve productivity, whereas our system is designed to create a sense of human connection among friends who happen to work in the same building, and explicitly *not* to support work. Second, CSCW systems tend to support awareness of physical presence and/or activities, while Affector supports awareness of *emotional* presence (mood) while through its distortions suppressing distracting and in this case privacy-intruding awareness of activity. As a research project, our greatest contribution is towards design for the complexity of human interpretation of the system, rather than optimal reproduction, representation and transmission of information by the system itself. In this respect, the system's philosophy is similar to that of eMoto, an open-ended system for communicating mood in email sent through mobile phones [5].

To address the complexity of interpretation, Affector is being built expressly for Simeon and Phoebe as a form of autobiographical design. This approach involves a careful examination and incorporation of the designers' subjective experiences in system design. We aim to consciously, self-reflexively, and responsibly design for ourselves, in the hopes that we will be able to create more rich and complex experiences than is possible when engineers attempt to take a more objective/external view of experience.

Implementation

Affector's implementation is inspired by Rodney Brooks's argument [2] that systems can appear intelligent and exhibit complex behavior without complex representation and manipulation of abstract information. Instead, Brooks's work is based on defining effective connections between sensors and effectors so that, when the system is placed in a complex environment, a complex sequence of actions is triggered which can be narrated as intelligent behavior.

Similarly, in Affector the output of the system's behavior may be readable as emotionally expressive without necessarily representing the emotions internally. A video screen is mounted on each side of the shared office wall to act as a virtual window. A video camera mounted under each screen captures images of the respective office occupant at work and transmits them to the neighboring office. On the way, the images are distorted in ways that may be read as representing emotion using visual algorithms developed by Eunyoung "Elie" Shin and Rev Guron, based directly on sensor readings in each office.

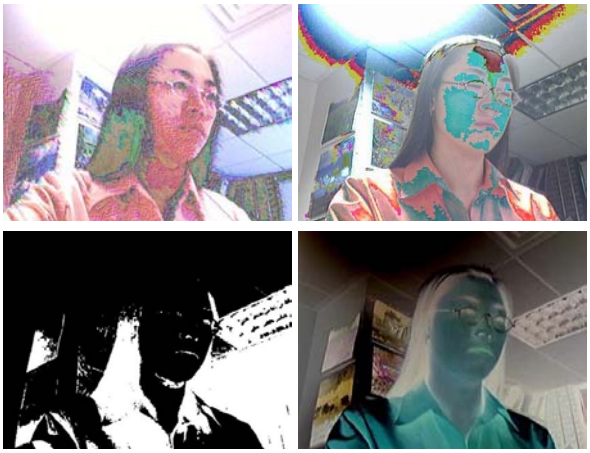

Figure 1: Example distortions produced by Affector.

The mapping between sensors (e.g. light levels in the office) to effectors (specific sets of distortions) is accomplished through a set of rules defined by the office occupants themselves. These rules select and combine visual distortions based on ambient information (currently visual attributes, in the future to be extended to include audio and potentially other sense modalities). Users of the system select and refine the rules until they seem, for them, to be accurately readable as expressing their friend's mood.

AFFECTOR: EVALUATION SKETCH

The Affector evaluation is a process of co-interpretation among the project participants, acting simultaneously as designers, as evaluators, and as users of the system. We are attempting to understand how the system influences our workaday lives, what effects we determine or discover to be desirable and why, and how the system is shaped to achieve these ends.

System Evaluation

The evaluation of the system itself is input for the critical and phenomenological evaluation of why the system was designed and how it is used. The key question in system evaluation for Affector is whether it works. Does the system in fact create opportunities for awareness of affective presence between two co-located people?

A conventional approach to system evaluation might first attempt to establish baseline measures of affective presence awareness and patterns of behavior without the system and then look for differences in these measures once the system is installed. For Affector, we will use three 'baseline' measures, not as conclusive evidence for whether the system works or not but as input for the evaluations discussed in the following sections.

The first baseline measure will establish patterns of face-to-face interaction between Phoebe and Simeon. It is first measured prior to installing the system; measurement continues through system use and after the system is removed. Physical counters have been placed in locations around the building, including each person's respective office, and common spaces such as the kitchen. Phoebe and Simeon use the counters to tally when a face-to-face interaction occurs and to indicate who initiated the interaction. At a surface level, if the patterns of face-to-face interaction change during the course of using Affector, it would suggest some type of effect, although it is difficult to say whether more or fewer interactions is an "improvement". More interesting for our purposes, is not the pattern change itself but how Phoebe and Simeon narrate this change - why did their patterns change and was this related to the system use?

The second baseline measure will begin once the system is installed (expected Feb. 2005). This measurement records how often each participant uses the system and how these patterns of usage change over time. As with the interaction measurement, an increase or maintenance in use versus a drop off would suggest success over failure. However, the actual numbers recorded are again less interesting than the way Phoebe and Simeon account for any change or non-change in frequency and sequencing of use.

Finally, the last baseline measure is defining the system's goals and use prior to installation, during installation, and after the installation is removed. Do the participants change their definition of what the system is about over time? This last measure and the stories told about the changing quantitative measures of system use and non-system interactions are input for the analysis of the design process and system use.

Critical Practice Evaluation

The second type of evaluation in examining Affector looks at the design choices made throughout the system's evolution. The very act of conceptualizing Affector is an act of critical practice because it proposes that 'affect' need not be reduced to a codifiable bit of information to represent and transfer. The conceptual design of Affector, therefore, challenges the predominant focus on systems geared toward making the communication of affect more efficient or harnessing affect for improved productivity. The effectiveness of this critical practice can be measured through its uptake in the design community for stimulating an alternative perspective to design for affect (e.g. perhaps through its acceptance to this workshop!).

However, the Affector is not only a speculative design but an actual designed system. The key areas being evaluated as a critical design practice involve understanding how Phoebe and Simeon evolve their design in response to their evolving definition of what they want the system to be: how they determine 'what works' as both designers and evaluators. This changing design practice will be evaluated two ways, 'internally' and 'externally.'

For the internal evaluation, they will draw from the stories developed during the evaluation of the system. In addition to measuring and narrating changes in pattern of use, we will also record what permutations, i.e. distortions, in the system are used. Phoebe and Simeon will record which permutations they feel are most effective, why, and how these choices change over time.

For the external evaluation, Tom Jenkins is working as a participant-observer to analyze the metaphors the project team uses to discuss and develop the project, applying and developing theories of the development of technology meaning as a social and cultural phenomenon from the social studies of science & technology.

Finally, the design process itself stands as an alternative to traditional design practices. In a traditional design process, designers attempt to create generic systems for generic users. In contrast, the design of Affector employs autobiographical design, designing for two very specific users. One of the questions for evaluation is whether this specific approach has scalability. This will be measured

through taking the final designed system and testing it with new pairs of users.

Evaluation of Use and Appropriation

In the final category of evaluation, we look at the use and appropriation of Affector as a lens for understanding how affect is constructed and communicated. This category of evaluation draws from the insights of the previous two. It uses the constructions of what works and why that were developed during the system evaluation and the insights from the critical practice evaluation to understand how the system is appropriated by its users.

Furthermore, evaluating the use of the system is similar to a psychological experiment where an experimental condition is imposed to explore constructs such as memory, identity, emotion, etc.. The implementation of Affector is an intervention, providing an experimental condition for understanding the phenomenon of affect. How might a technological system for affect influence the way we experience and define affect? The fact that many systems for affect do depend on discrete, transferable representations of affect sheds light not only on system design, but on how we conceptualize affect. By creating a system that does not depend on internal representations, we can also evaluate new ways of conceptualizing the expression of affect.

This form of evaluation will depend upon a phenomenological analysis of how the system is appropriated. How does its use vary from the original designs? How do Phoebe and Simeon develop a common language or common understanding around what the system means?

The distortions that Phoebe and Simeon develop initially exist as iconic demonstrations of affect. The distortions do not begin, as is typical in affective computing systems, as re-mappings of words to visual effects, e.g. a red color wash means 'I'm busy'; rather, the distortions are a visual impression of some indication of activity. Eventually, however, the distortions may become symbolic of certain meanings. By asking Phoebe and Simeon at multiple times throughout the installation to interpret visual displays, we can see how their interpretations evolve, what resources they draw from to build these interpretations, how consistent they are, and how resonant their interpretations are with each other.

In the critical practice evaluation, we look at how definitions, experiences, and expectations of both affect and technology inform the design of Affector. In this third category of evaluation, we look at the appropriation of the system by its users and how Affector in use informs our conceptualizations of affect and technology.

CONCLUSION AND NEXT STEPS

Our major challenge in evaluating Affector is in combining these three forms of evaluation, which are not necessarily compatible. System evaluation normally focuses fairly exclusively on system functionality, placing blame and credit for any successes or failures on the system itself. Evaluation of use and appropriation takes the opposite tack, looking at users as active appropriators of system functionality and placing blame or success in the context of system use, rather than in the system itself. Critical practice evaluation normally takes place independently, aiming a suspicious eye at unconscious assumptions and values that influence the design practice. Yet we believe these three forms of evaluation are all essential to understanding whether, how, why, and in what sense Affector works, and we hope through our practice and through our participation in this workshop to develop strategies for combining these three for a more holistic approach to affective system evaluation.

ACKNOWLEDGMENTS

Thanks to our collaborators Eunyoung "Elie" Shin, Yevgeniy "Eugene" Medynskiy, and Rev Guron. This project is funded in part by NSF Grant 0238132. We are grateful for inspiration and support by our Affective Presence partners, Bill Gaver, Geri Gay, Kristina Höök, Michael Mateas, Carol Minton Morris, and the Intel People & Practices group.

REFERENCES

1. Boehner, Kirsten , Mo Chen, and Zheng Liu. The Vibe Reflector: An Emergent Impression of Collective Experience. CHI 2003 Workshop on Providing Elegant Peripheral Awareness. http://research.microsoft.com/~jjcadiz/CHI2003Workshop/submissions/TheVibeReflector-CHI.pdf
2. Brooks, Rodney A. Intelligence without representation, Artificial Intelligence, v.47 n.1-3, p.139-159, Jan. 1991
3. Höök, Kristina, Phoebe Sengers, and Gerd Andersson. "Sense and Sensibility: Evaluation and Interactive Art." In *2003 Conference on Computer-Human Interaction (CHI)*, 2003.
4. Sengers, Phoebe, Rainer Liesendahl, Werner Magar, Christoph Seibert, Boris Müller, Thorsten Joachims, Weidong Geng, Pai Martensson, and Kristina Höök. "The Enigmatics of Affect." *Conference on Designing Interactive Systems (DIS)*. London, England, June 2002.
5. Sundström, P., Ståhl, A., and Höök, K. eMoto – Affectively Involving both Body and Mind, to appear, Proceedings of CHI '05, Portland Oregon, USA, 2005.