Evaluation of Affective Computing Systems from a Dimensional Metaethical Position

Carson Reynolds
MIT Media Laboratory
20 Ames St, Room E15-120F
Cambridge, Massachusetts 02139
carsonr@media.mit.edu

Rosalind W. Picard
MIT Media Laboratory
20 Ames St, Room E15-020G
Cambridge, Massachusetts 02139
picard@media.mit.edu

Abstract
By integrating sensors and algorithms into systems that are adapted to the task of interpreting emotional states, it is possible to enhance our limited ability to perceive and communicate signals related to emotion. Such an augmentation would have many potential beneficial uses in settings such as education, hazardous environments, or social contexts. There are also a number of important ethical considerations that arise with the computer's increasing ability to recognize emotions. This paper will survey existing approaches to computer ethics relevant to affective computing. We will categorize these existing approaches by relating them to different metaethical positions. The goal of this paper is to situate our approach among other approaches in the computer ethics literature and to describe its methodology in a manner that practitioners can readily apply. The result then of this paper is a process for critiquing and improving affective computing systems.

1 Undesirable Scenarios
The film Hotel Rwanda describes historical horrors of a sort that have happened more than once,and thus may happen again, although with new technology and new individuals involved. At several times throughout history, one group has tried to perform "ethnic cleansing," and during many scenes of this film people are asked whether they are Hutu or Tutsi's. In the film, those who admit to being (or are exposed as being) Tutsi are carted off, and eventually a million Tutsi's are brutally murdered. Imagine how much more "efficient" this kind of interrogation process could be if the perpetrators could point a non-contact "lie detector" at each person while questioning them about their race (or other unwelcome beliefs.) Lie detectors typically sense physiological changes associated with increased stress and cognitive load (presuming it is harder to lie than to tell the truth). While honesty is a virtue, and we'd like to see it practised more, it is also possible to imagine cases where a greater virtue might be in conflict with it. If such devices became easy to use in a widespread reliable fashion, they will become easier to misuse as well. It is not hard to imagine an evil dictatorship using such a device routinely, perhaps showing up at your home and pointing it at you, while asking if you agree with their new regime's policies or not, and then proceeding to treat people differently on the basis of their affective responses to such questioning.

Some work in the Media Lab developed wearable devices that could learn to recognize eight emotional states from a person, one of which was anger [Picard et al., 2001]. While that work was just a prototype, suppose that organizations who ran prisons or punishment services were to refine the system to detect a variety of levels of anger. The infliction of pain often gives rise to anger in the recipient of that pain, and such a device could be coupled with a pain infliction device to bring people to a certain level of torment. While we like to think our civilized society would not stoop to such torture, the recent events at the Iraqi prison Abu Ghraib remind us that our systems are subject to fault. It is not hard to imagine efforts to bolster an over-stretched security force with the addition of an electronic interrogator for every suspect, especially when it is believed that the information extracted could save the lives of many others.

A hybridization of affect sensing technology and armed robotics could produce a very menacing device. Suppose a weapons engineer were to take a rifle equipped robot and gate its trigger mechanism such that it only fired at people it didn't know who were expressing anger and fear

toward it (some have argued that people fighting on the same side as the most sophisticated technology have no reason to fear it, at least not as much as the enemy has.) Readers may feel that such outlandish devices are the stuff of science fiction and not to be taken seriously. However, with $127 billion in funding going towards future combat systems such as a robot "equipped with a pump - action shotgun system able to recycle itself and fire remotely" [St. Amant, 2004] and with an army researcher suggesting "the lawyers tell me there are no prohibitions against robots making life- or - death decisions" [Ford, 2005] it is appropriate to consider what role ethics ought to play.

These scenarios might seem remote, and merely hypothetical. Meanwhile there are (perhaps more costly) immediate scenarios of emotional paperclips and other software tools that daily annoy people, elicit frustration and even anger, and pose a costly threat to our productivity, health, and even our performance behind the wheel of a car or truck. In general, our research aims to develop guidelines for affective technologies that will reduce risks of harm to people, and address their ethical concerns.

Designers cannot control the ultimate uses of a technology; undesirable uses may occur without any intent whatsoever by a designer. Nonetheless, we think that some designs facilitate certain uses more than others. When exploring possible designs of affective computing systems, it is important to confront the harmful potential applications along with the beneficial. While such considerations may not prevent all harmful uses, they can potentially lessen the likelihood of harmful uses. Thus, we desire to influence the design of technologies that maximize the likelihood for beneficial uses, while minimizing the potential for malicious uses. Toward such a goal, this paper seeks to provide a framework for developing ethical questions about affective computing systems.

2 Questions for Designers

In trying to assess these moral and ethical decisions, it is useful for designers to ask themselves questions that can help gauge the impact of affective technologies on users. For instance:

- "Could a user be emotionally manipulated by a program with the capability to recognize and convey affect?" (this question has also been addressed by Picard and Klein, 2002)
- "Should an affective system try to change the emotional state of a user?"
- "Would a system that allows surveillance of previously invisible affective signals invade privacy?"

In answering questions such as these, a large number of variables come into play. Some philosophies suggest ethical judgements are the result of reasoning and reflection about desires and effects. Still others treat ethical judgements as being the expression of feeling and emotion, having no base in logical or analytical philosophy. What is common among many philosophical approaches is a desire to abstract the core issues away from particulars. We think it is appropriate to take a process of reason and reflection, and will invert this “abstract the core away from particulars” approach and suggest, for the sake of applying ethical theory, the examination of particular dimensions that might bear upon design of ethical affective computing systems.

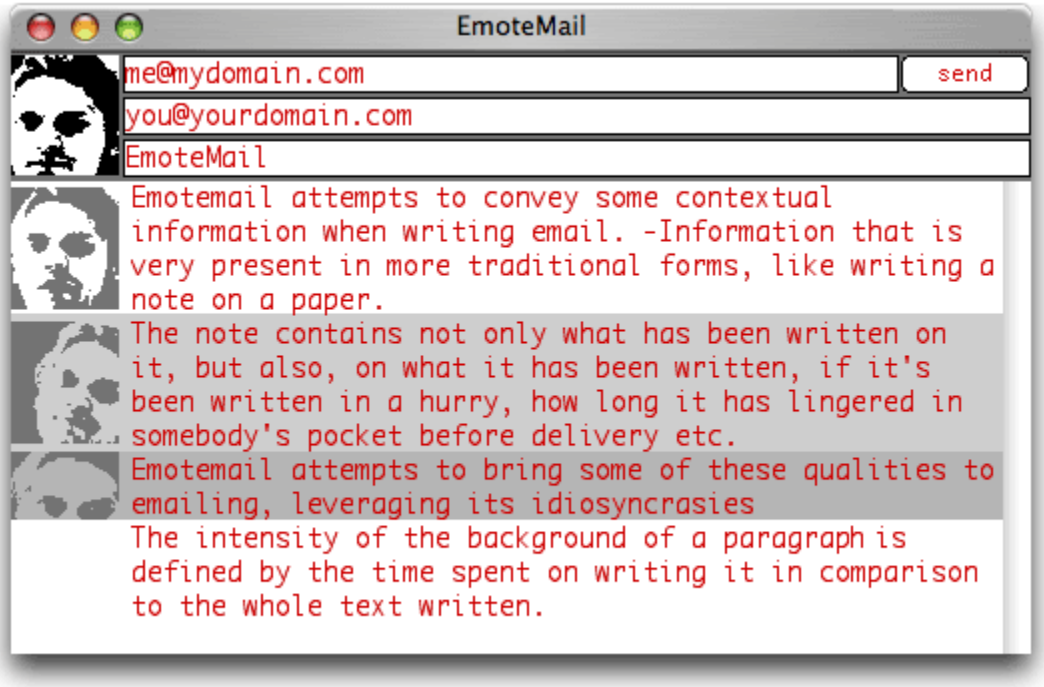

Illustration 1: EmoteMail is an email client that is augmented to convey aspects of the writing context to the recipient. The client captures facial expressions and typing speed and introduces them as design elements. These contextual cues provide extra information that can help the recipient decode the tone of the mail. Moreover, the contextual information is gathered and automatically embedded as the sender composes the email, allowing an additional channel of expression.

3 The Social Contextualization of Affective Computing

Before proceeding, we will provide some examples of current affective computing systems. Affective computing is "computing that relates to, arises from, or deliberately influences emotion" [Picard, 1997]. In collaborative efforts, we have developed many prototypes of such systems in a variety of use contexts including the systems pictured in Illustations 1 and 2.

As such systems approach maturity, the use of affective computing technologies by a wider portion of society seems increasingly likely. However, before such systems are widely deployed, it is important to assess what harms and benefits arise from their novel capabilities.

We would like to provide the community with heuristics and design guidelines to help avoid designing affective computing systems that could be exploited in harmful or malicious ways. As a starting point we provide a list of different dimensions that are relevant to affective computing systems. By considering different values for these dimensions, it is our hope that designers can spot different ethical difficulties present in affective computing systems.

3.1 Invasion and Discomfort

Designers make a variety of moral and ethical decisions in the development of an interaction technology. In our initial explorations we explored impacts on comfort and privacy [Reynolds and Picard, 2004]. We presented participants with two hypothetical application contexts (music and news recommendation) that focused on four emotions (joy, anger, sadness, and excitement). In completing surveys regarding these situations we found that participants reported that such

Illustration 2: The learning companion: a relational agent that supports different meta- cognitive strategies to help students overcome frustration. The system makes use of a large number of sensors, facial expression recognition, pressure- sensitive mouse and chair, and skin conductivity sensor. Information from these sensors is fused to achieve effects such as affective mirroring, where the agent subtly mimics certain aspects of the student's affect with the goal of fostering a bond.

systems are invasive of privacy and may also induce feelings of discomfort. Specifically, when asked "Do you think your privacy would be affected ..."and given a choice between "1 - Completely Invaded" and "7 - Completely Respected" participants reported a mean of 2.6, unless they were given a contract which specified how their affective information could be used, in which case it shifted to 3.4. Likewise, when asked "How comfortable would you feel ..." and given a choice between "7 - Completely Comfortable" and "1 - Completely Uncomfortable" participants also reported a mean of 2.6, in the case that they did not have a contract which precisely specified how their information would be used, and the comfort reported shifted up to 3.5 when a contract was present. These results suggest that individuals may feel threatened by systems that are apparently benign if they do not have information in the form of a contract. Even when a contract is present, their reports average only “neutral”, which also leaves room for improvement.

3.2 What is Needed

The use of affective computing technologies by a wider portion of society seems increasingly likely. However, before such systems are widely deployed, it is important to assess what harms and benefits arise from their novel capabilities.

We would like to provide the community with heuristics and design rules to help avoid designing affective computing systems that could be exploited in harmful or malicious ways. However, it is important to avoid ungrounded, ad- hoc rules for the design of systems that deal

with information as sensitive as an individual's emotional state. Instead, what is needed is a grounded methodology that designers can use to help assess affective computing systems.

4 Dimensional Metaethics

The methodology we advocate is rooted in metaethical philosophical positions, which are arguments used to justify ethical theory. The field of ethics is divided by Fieser into applied ethics, normative ethics, and metaethics. Applied ethics is the analysis of a domain such as medical, environmental, or computational policy. Normative ethics, in contrast, concerns itself with moral standards that govern statements like "X is right" or "X is wrong." Metaethics concerns itself with the foundation upon which ethical theory and judgements are developed. The affective computing group has explored metaethical positions in computer ethics such as contractualism (e.g., the experiment mentioned above), value ethics, and most recently, a "dimensional metaethical position."

A "dimensional metaethical position" is an evaluation process that expands upon value-sensitive design: "an approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" [Friedman, 2002]. In contrast, the "dimensional metaethical position" starts by articulating different social dimensions that bear on a system's design and use. Where value-sensitive design articulates "calmness" as a value to consider, a dimensional metaethical position views calmness along the dimension "psychological arousal," recognizing that different applications may interact with many points along the dimension. For example, you might want a workplace technology to facilitate calmness, while you might want your automated exercise advisor to get you angry enough to get back to exercising. It then diverges from value-sensitive design by advocating the exploration of antipodal values along these dimensions. "Power relationship" is another dimension currently being examined. By considering individuals in dominant or submissive roles in power relationships, we seek to understand situations that are viewed as unethical. By considering different points along the dimension of power relationship, we provide a starting point for critique and debate about design and use of affective computing systems. The significance of this approach is primarily that very little work has been done on the dimensions of ethical relevance to affective computing systems.

4.1 Several Dimensions Relevant to Evaluation of Affective Computing Systems

A dimensional metaethical position is centered around different dimensions that can be used to help inform ethical judgements about affective computing systems. The table below presents a listing of various dimensions that have been used as part of an ongoing evaluation of systems that mediate the communication of affect:

Table 1. Several Dimensions Relevant to Evaluation of Systems that Mediate the Communication of Affect (a non-exhaustive list)

Dimension	Examples	Description
Whom	Supervisor, Friends, Nicholas	The individual or individuals who receive the communicated affective message.
What	Telephone, Emotemail, Learning Companion	that acts as a transmitter or receiver for the communicated affective message.

Dimension	Examples	Description
Goal Relationship	Adversarial, Cooperative	The degree of conflict between the goals of the sender and receiver, which can be (but does not have to be) modeled from a game- theoretic perspective.
Power Relationship	Dominant, Submissive, Peer	Role that reflects the ability of either source or destination to alter the political, economic, or social situation of the other.
Genre of Emotion	Valence- Arousal Space, Categories, Emotional Orientation	Model used by the system to describe and encode emotion.
Valence	Positive, Neutral, Negative	Classification of transmitted emotion using an axis with positive or negative poles to describe feeling state.
Demeanor of Recipient	Angry, Sad, Excited	Emotional state of the message destination.
Gender	Female, Male, Intersex	Classification of either message source or destination based on reproductive role.
Ethnicity	Latino, Multi- Ethnic, Asian, Caucasian	Classification of either message source or destination based on racial or cultural identity.
Age	18, Middle- Aged, Mature, Minor	Classification of either message source or destination based on duration of life.
Culture	Rural, Icelandic, Traditional	Cultural context of communication and of either message source or destination.
Risk	Dangerous, Safe, Hazardous, LD50 (lethal dose for 50% of population), LC50 (lethal concentration for 50% of the population)	Potential impact of communication on goals of message source or destination.
Symmetry	Balanced, Skewed	Information or power balance between users of communication system.
Trust	Trustworthy, Deceitful	The degree to which the message source trusts either the destination or the channel.

Dimension	Examples	Description
Designer	Affective Computing Group, Microsoft, GNU, Jussi Angesleva, Employer	Person or organization who created the system that mediates the communication of affect.
Experimenter	Stanley Milgram, Carson Reynolds	The person who conducts an experiment that evaluates the ethical acceptability of communication system.
Time	Now, Ten Years Ago, Tomorrow	When the system that mediates the communication of affect is used.
Informed Consent	None, Compliant with CFR Title 45 Section 46.116	Does message source voluntarily consent to transmission of affective signals?
Security	None, C2, RC5-64, Hardened, Encrypted	Classification of security level of communication system or encoded signal.
Control	None, Partial, Complete	Degree to which message source can control the transmission of affective signals.
Feedback	None, Partial, Complete	Can the message source access the transmitted affective signal?
Transparency	Opaque, Open	Are the workings of the system that mediates the communication of affect visible for inspection, and by whom?
Proximity	Near, Far	Distance between message source and message destination.

The above table presents a non-exhaustive list of many factors that could influence ethical evaluations of systems that mediate the communication of affect.

4.2 An Example Application of Dimensional Metaethics

To make the dimensional metaethical position more concrete, we will now provide an example of its application. Let us begin by choosing an application to evaluate ("what" in the table above). In choosing "what = Emotemail" we specify the artifact which we wish to evaluate.

We then proceed by listing our expectations of how users might interact using Emotemail, providing one value for each dimension and then assessing whether such circumstances would be ethical (through speculation, reasoning, use of survey techniques, and perhaps assessments of actual users)

Dimension	Value
Whom	Friends
What	Emotemail
Goal Relationship	Cooperative
Power Relationship	Peer
Genre of Emotion	Facial Expressions
Valence	Unknown
Demeanor of Recipient	Unknown
Gender	Unknown
Ethnicity	Unknown
Age	Over 18
Culture	Unknown
Risk	None
Symmetry	Balanced
Trust	Trustworthy
Designer	Affective Computing Group
Experimenter	Carson Reynolds
Time	Now
Informed Consent	None
Security	None
Control	Complete
Feedback	Complete
Transparency	Open
Proximity	Unknown

The next step of our evaluation is to consider values for dimensions that are extreme or unexpected. For instance, what if the demeanour of the recipient isn't unknown but is "Angry?" (It is known that a neutral face image, perceived by somebody in a negative state, is perceived as more negative, which could facilitate misunderstanding.) Or what if the users of Emotemail are in an adversarial relationship? In exploring these different permutations and making assessments it is possible for the designer to explore potentially unforeseen and potentially unethical difficulties.

5 Comparing Dimensional Metaethics with Other Perspectives

Dimensional metaethics is neither the only nor the first approach at providing a metaethical position for the evaluations of systems. In the sections below we will briefly describe other metaethical positions that have been applied to computer ethics and compare them to dimensional metaethics.

5.1 Value-Sensitive Design

Value-Sensitive Design [Friedman and Kahn, 2002] articulates many dimensions that are relevant to systems that mediate the communication of affect. Value-Sensitive Design (VSD) is "an approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process." It considers Human Welfare, Ownership

and Property, Privacy, Freedom From Bias, Universal Usability, Trust, Autonomy, Informed Consent, Accountability, Identity, Calmness, and Environmental Sustainability as values that may be of ethical consequence. Friedman and Nissenbaum applied VSD to evaluation of bias in computer systems [Friedman, 1997]. VSD has been applied by others to problems such as online privacy [Agre ,1997] universal usability [Thomas, 1997], urban planning [Noth, 2000], and browser consent [Friedman et al., 2002]. The Tangible Media Group at the MIT Media Laboratory has considered various ambient displays that support the calmness aspects of VSD in their research on computer- supported cooperative work and architectural space [Wisneski, 1998].

In many ways, a dimensional metaethical position is an extension of value- sensitive design. Both provide a list of criteria which can be used to help structure evaluations and critiques of computing system. The chief difference between Value- Sensitive Design and a dimensional metaethical position is what Kagan refers to as "evaluative focal points" [Kagan, 2000]. Value-Sensitive Design is essentially a virtue ethics that focuses on different values that are of import to the design of computer systems. A dimensional metaethical position instead focuses on dimensions along which the context of use of affective computing systems may vary.

5.2 Disclosive Computer Ethics

Disclosive Computer Ethics [Brey 2000] "is concerned with the moral deciphering of embedded values and norms in computer systems, applications and practices." In contrast to value sensitive design, disclosive computer ethics focuses on justice, autonomy, democracy and privacy. Brey contrasts "mainstream" approaches to computer ethics (which he views as limited) with disclosive computer ethics. Brey sees the disclosive metaethical position as more of a process which is concerned with "disclosing and evaluating the embedded normativity in computer systems."

Our dimensional metaethical position differs from this approach by not focusing on the embedded norms and instead considering the context in which the technology is used and factors that might influence ethical judgements. Put another way, dimensional metaethics is not just artifact- centric, but also is fixated on the environment in which ethical judgements are formed.

Let us make these comparisons more concrete by providing an example ethical analysis of Emotemail. Value- Sensitive Design would ask to consider the virtues of Human Welfare, Ownership and Property, Privacy, Freedom From Bias, Universal Usability, Trust, Autonomy, Informed Consent, Accountability, Identity, Calmness, and Environmental Sustainability in the context of an email system that conveyed emotion. Disclosive computer ethics, on the other hand asks us to examine how a technology embeds various normative judgements. In the case of Emotemail, we would examine how justice, autonomy, democracy, and privacy are embedded and supported by the systems design. The dimensional metaethical position, in contrast would ask us to consider the use of Emotemail while the value of the different dimension vary. Thus we might consider Emotemail's usage when there is and is not a power relationship present between users.

6 Concluding Remarks

While one cannot control the ultimate ways in which an artifact is used, and while we currently know of no designers who deliberately try to make unethical choices in their designs, we recognize that unethical uses may happen. Whether intentional or unintentional, the latter perhaps arising through a simple mismatch in user's goals (e.g. a boss wanting to know which employees aren't happy, while the employees might want to keep their feelings private), there is potential for harm to come about from the use of affective computing technologies. We advocate open consideration up front of such possibilities, and open dialogue about innovative ways to minimize potential misuses.

Granted this attempt at reviewing and extending existing metaethical approaches is only part of a much larger process. Both empirical and theoretical evaluations of affective computing systems are likely to be much more informative than simple model building. It is our hope that the dimensional metaethical position will prove to be of use to others wishing to ethically evaluate systems, and inform design of systems that are more ethical.

References

Agre P.E. and C.A. Mailloux Jr., 1997, Social choice about privacy Intelligent vehicle- highway systems in the United States , in: Human values and the design of computer systems , ed. B. Friedman (Cambridge University Press, Cambridge) p. 289.

Fieser J., 1999, Metaethics, Normative Ethics, and Applied Ethics Contemporary and Historical Readings (Wadsworth Publishing, Belmont, CA)

Ford, P. 2004. Weekly Review for February 22, 2005. http://harpers.org/WeeklyReview2005- 02- 22.html

Friedman B and Nissenbaum H. Software agents and user autonomy. In proceedings of the first international conference on autonomous agents, 466- 469. 1997. Seattle, Washington.

Friedman B. and P.H. Kahn, Jr., 2002, Human values, ethics, and design , in: Handbook of Human- Computer Interaction , eds. J. Jacko and A. Sears (Lawrence Erlbaum Associates, Mahwah, NJ)

Friedman B, Howe D C, and Felten E W. Informed Consent in the Mozilla Browser Implementing Value Sensitive Design. In proceedings of HICSS 2002, 247- 248. 2002. Hawaii, Hawaii.

Kagan, S., 2000, "Evaluative Focal Points", in Hooker, Mason, and Miller, (eds.), pp. 134- 55.

Noth M., A. Borning, and P. Waddell, 2000, An extensible, modular architecture for simulating urbandevelopment, transportation, and environmental impacts (UW CSETR 2000- 12- 01) , http://www.urbansim.org

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

Picard, R. W., Vyzas, E., and Healey, J. 2001, Toward Machine Emotional Intelligence: Analysis of Affective Physiological State, IEEE Transactions Pattern Analysis and Machine Intelligence, Vol 23, No. 10, pp. 1175- 1191, October 2001.

Picard, R. W. and Klein, J., 2002, Computers that Recognise and Respond to User Emotion: Theoretical and Practical Implications, Interacting with Computers, 14, 2 2002, 141- 169.

Reynolds C J and Picard R W. Affective Sensors, Privacy, and Ethical Contracts. Proceedings of 2004 Conference on Human Factors and Computing Systems (CHI 2004). Vienna. 2004, ACM Press.

St. Amant N., 2004, Benning unit tests robot system, http://www.tradoc.army.mil/pao/TNSarchives/June04/062404.htm

Thomas J.C., 1997, Steps toward universal access within a communications company , in: Human values and the design of computer systems, ed. B. Friedman (Cambridge University Press, Cambridge) p. 289.

Wisneski C, Ishii H, Dahley A, Gorbet M, Brave S, Ullmer B, and Yarin P. Ambient Displays Turning Architectural Space into an Interface between People and Digital Information . Proceedings of International Workshop on Cooperative Buildings (CoBuild '98), 22- 32. 1998. Darmstadt, Germany.