M.I.T Media Laboratory Technical Report #534

Chapter to appear in: "Emotions in Humans and Artifacts," ed. by R. Trappl, P. Petta and S. Payr.

What does it mean for a computer to “have” emotions?

Rosalind W. Picard
MIT Media Laboratory
E15-392, 20 Ames St., Cambridge MA 02139
http://www.media.mit.edu/~picard

Introduction

There is a lot of talk about giving machines emotions, some of it fluff. Recently at a large technical meeting, a researcher stood up and talked of how a Barney stuffed animal (the purple dinosaur for kids) “has emotions.” He did not define what he meant by this, but after repeating it several times, it became apparent that children attributed emotions to Barney, and that Barney had deliberately expressive behaviors that would encourage the kids to think Barney had emotions. But kids have attributed emotions to dolls and stuffed animals for as long as we know; and most of my technical colleagues would agree that such toys have never had and still do not have emotions. What is different now, which prompts a researcher to make such a claim? Is the computational plush an example of a computer that really does have emotions?

If not Barney, then what would be an example of a computational system that has emotions? I am not a philosopher, and this paper will not be a discussion of the meaning of this question in any philosophical sense. However, as an engineer I am interested in what capabilities I would require a machine to have before I would say that it “has emotions,” if that is even possible.

Theorists still grapple with the problem of defining emotion, after many decades of discussion, and no clean definition looks likely to emerge. Even without a precise definition, one can still begin to say concrete things about certain components of emotion, at least based on what is known about human and animal emotions. Of course much is still unknown about human emotions, so we are nowhere near being able to model them, much less duplicate all their functions in machines. Also, all scientific findings are subject to revision – history has certainly taught us humility, that what scientists believed to be true at one point has often been changed at a later date.

I wish to begin by mentioning four motivations for giving machines certain emotional abilities (and there are more). One goal is to build robots and synthetic characters that can emulate living humans and animals – for example, to build a humanoid robot. A second goal is to make machines that are intelligent, even though it is also impossible to find a widely accepted definition of machine intelligence. A third goal is to try to understand human emotions by modeling them. Although I find these three goals intriguing, my main focus is on a fourth: making machines less frustrating to interact with. Toward this goal, my research assistants and I have begun to develop computers that can identify and recognize situations that frustrate the user, perceiving not only the user’s behavior and expressions, but also what the system was doing at the time. Such signs of frustration can then be associated with potential causes for which the machine might be responsible or able to help, and the machine can then try to learn how to adjust its behavior to help reduce frustration.   It may be as simple as the computer noticing that lots of fancy “smart” features are irritating to the user, and offering the user a way to remove all of them. Or, it may be that the computer’s sensitive acknowledgment of and adaptation to user frustration simply leads to more productive and pleasing interactions. One of the key ideas is that the system could associate expressions of users, such as pleasure and displeasure, with its own behavior, as a kind of reward and punishment.   In this age of adaptive, learning computer systems, such feedback happens to be easy and natural for users to provide.

My first goal thus involves sensing and recognizing patterns of emotional information - dynamic expressive spatial-temporal forms that influence the face, voice, posture, and ways the person moves - as well as sensing and reasoning about other situational variables, such as if the person re-typed the same word many times and is now using negative language. All of this is what I refer to in shorthand as “recognizing emotion, “ although I should be clear that it means the first sentence of this paragraph, and not that a computer can know your innermost emotions, which involve thoughts and feelings that no person besides you can sense. But once a computer has recognized emotion, what should it do? Here lies my second main goal: giving the computer the ability to adapt to the emotional feedback in a way that does not further frustrate the user. Although “having emotion” may help with the first goal, I can imagine how to achieve the first goal without this ability. However, the second goal involves intricacies in regulating and managing ongoing perceptual information, attention, decision-making, and learning. All of these functions in humans apparently involve emotion. This does not mean that we could not possibly implement them in machines without emotion. At the same time, it appears to be the case that all living intelligent systems have emotion in some form, and that humans have the most sophisticated emotion systems of all, as evinced not just by a greater development of limbic and cortical structures, but also by greater facial musculature, a hairless face, and the use of artistic expression, including music, for expressing emotions beyond verbal articulation.

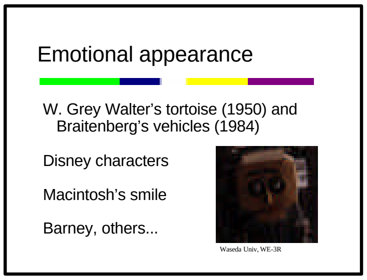
Part of me would love to give a computer the ability to recognize and deal with frustration as well as a person can, without giving it emotions.   I have no longing to make a computer into a companion; I am quite content with it as a tool. However, it has become a very complex adaptive tool that frustrates so many people that I think it’s time to look at how it can do a better job of adapting to people. I think emotion will play a key role in this. Let’s look more closely at four components of emotion that people have, and how these might or might not become a part of a machine.

A computer that “has emotions,” in the sense that a person does, will be capable of:

1. Emotional appearance
2. Multi-level emotion generation
3. Emotional experience
4. Mind-body interactions

Components of emotion

I find it useful to identify at least four components when talking about emotions in the context of what one might want to try to implement in machines. Some of these components already exist in some computational systems. The components are 1. Emotional appearance, 2. Multiple levels of emotion generation, 3. Emotional experience, and 4. (A large category of) Mind-body interactions. These four components are not intended to be self-evident from their short names, nor are they intended to be mutually exclusive or collectively exhaustive. Let me say what I mean by each, and why all four are important to consider.



Emotional appearance

W. Grey Walter's tortoise (1950) and Braitenberg's vehicles (1984)

Disney characters

Macintosh's smile

Barney, others...

Waseda Univ, WE-3R

Emotional appearance

Barney the stuffed animal sometimes *sounds as if* he is happy. Like a 3-D animated cartoon, he has expressions and behaviors that were designed to communicate certain emotions. "Emotional appearance" includes behavior or expressions that *give the appearance that* the system has emotions.

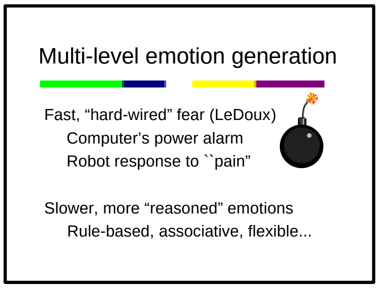
This component is the weakest of the four, in the sense that it is the easiest of the four to produce, at least at a superficial level. However, I include it because this quality is all that an outside observer (non-designer of the system, who cannot access or decipher its inward functions) has at his or her disposal in order to judge the emotional nature of the system. By and large, it is what the crew in the film *2001: A Space Odyssey* did not perceive about the computer HAL until the end of the film, otherwise they might have obtained earlier clues about HAL's increasingly harmful emotional state, which at the end of the film is illuminated when HAL finally says, "I'm afraid, Dave, I'm afraid." This component is also the most commonly implemented in machines today – primarily in agents and robots that display emotional behaviors in order to "look natural" or to "look believable."

Because emotional appearance results largely from emotional behavior, and because I include the making of facial, vocal, and other expressions as kinds of behavior, I have previously referred to this component as "Emotional Behavior." I am here changing my two-word description since a couple colleagues at the Vienna workshop argued that it was confusing; however, I am not changing what it refers to, which remains the *emotional appearance of the system's behavior*.

Examples of systems with behaviors that appear to be emotional include the tortoises of W. Gray Walter (1950) and Braitenberg's Vehicles (Braitenberg 1984). When one of Braitenberg's little vehicles approached a light or backed rapidly away from it, observers described the behavior as "liking lights" or as "acting afraid of lights," both of which involve emotional attribution, despite that the vehicles had no deliberately designed internal mechanisms of emotion. Today there are a number of efforts to give computers facial expressions; the Macintosh has been displaying a smile at people for years, and there is a growing tendency to build animated agents and other synthetic characters and avatars that would have emotional expressions. These expressive behaviors may result in people saying the system is "happy" or otherwise, because it appears that way.

I think all of us would agree that the examples just given do not have internal feelings, and their behavior is not generated by emotions in the same sense that human or animal behavior is. However, the boundary is quickly blurred: Contrast a machine like the Apple Macintosh, which shows a smile because it is hardwired to do that in a particular machine state, and a new "emotional robot," which shows a smile (Johnstone, 1999) because it has appraised its present state as good and its present situation as one where smiling can communicate something useful. The Mac's expression signals that the boot-up has succeeded and the machine is in a satisfactory state for the user to proceed. However, most of us would *not* say that the Mac is happy. More might say that the robot is happy, in a rudimentary kind of way. But, if the robot's happy facial expression were driven by a simple internal state labeled "satisfaction," then it would really be no different than the Mac's display of a smile. As the generation mechanisms become more complex and adapted for many such states and expressions, then the argument that the expression or behavior really arose from an emotion becomes more compelling. The more complex the system, and the higher the users expectations, the harder it also becomes for the system's designer to craft the appearance of natural, believable emotions. Nonetheless, we should not let mere complexity fool us into thinking emotions are there.

If a system really has emotions, then we expect to see those emotions influence and give rise to behavior on many levels. There are the obvious expressions and other observable emotional behaviors, like saying "Humph," and turning abruptly away from the speaker; however, emotions also modulate non-emotional behaviors: the way you pick up a pen (a neutral behavior) is different when you are seething with anger vs. when you are bubbling with delight. True emotions influence a number of internal functions, which are generally not apparent to anyone but the designer of the system (and in part to the system, to the extent that it is given a kind of "conscious awareness" of such.) Some of emotion's most important functions are those that are unseen, or at least very hard to see.   The body-mind mechanisms for signaling and linking the many seen and unseen functions are primarily captured by the fourth component, which I'll describe shortly.



Multi-level emotion generation

- Fast, “hard-wired” fear (LeDoux)
  - Computer’s power alarm
  - Robot response to ``pain”
- Slower, more “reasoned” emotions
  - Rule-based, associative, flexible...

Multiple Levels of Emotion Generation

Animals and people have fast subconscious brain mechanisms that perform high-priority survival-related functions, such as the response of fear in the face of danger or threat. LeDoux (1996) has described the sub-cortical pathway of fear’s “quick and dirty” mechanism, which precedes cortical involvement. This level of pre-conscious, largely innate, but not highly accurate emotion generation appears to be critical for survival in living systems. One can imagine giving robots and machines sensors that operate at a similar level – in a relatively hard-wired way, detecting when the system’s critical parameters are in a danger zone, and triggering rapid protective responses, which can shortly thereafter be modified by slower more accurate mechanisms.

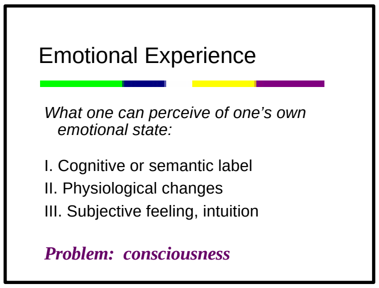
The level of emotions just described stands in contrast with slightly slower (although still very fast) emotion generation that tends to involve higher cortical functions and may or may not involve conscious appraisals. If you jump out of the way of a snake, and suddenly realize it was only a stick, then that was probably an instance of the fast subconscious fear generation mechanism. In contrast, if you hear that a convicted killer has escaped a nearby prison, and consequently decide that you don’t want to leave the house, then it is likely that your thoughts generated a form of a learned fear response, which subsequently influenced your decision. You may have never seen a convicted killer, but you cognitively know that such a person could be dangerous, and you associate with it a response that you learned from a similar but real experience. This learned fear response engages some of the same parts of the brain as the lower-level quick version of fear, but it additionally involves reasoning and cortical appraisal of an emotional situation.

Some of the most common methods of “implementing emotions” in computers involve constructing rules for appraising a situation, which then give rise to an emotion appropriate to that situation. An example is the OCC model (Ortony et al, 1988), which was not designed to synthesize emotions, but rather to reason about them, but works in part for either. Consider the generation of joy, which involves deciding that if an event happens, and that event is desirable, then it may result in joy for oneself or in happiness for another. A machine can use this rule-based reasoning either to try to infer another’s emotion, or to synthesize an internal emotional state label for itself. All of this can happen in the machine in a cold and logical way, without anything that an outsider might observe as emotion. It can happen without any so-called conscious awareness or “feeling” of what the machine is doing. This kind of “emotion generation” does not need to give rise to component one – emotional appearance – or to the other two components listed below, but it could potentially give rise to all of them. In a healthy human, such emotional appraisals are also influenced by ones feelings, via many levels of mechanisms.

People appear to be able to reason in a cold way about emotions, with minimal if any engaging of observable bodily responses. However, more often there seem to be bodily changes and *feelings* associated with having an emotion, especially if the emotion is intense. An exception arises in certain neurologically impaired patients, e.g., see accounts in (Damasio, 1994), that show minimal signs of such somatic concomitants of emotion. If you show these patients grotesque blood-and-guts mutilation scenes, which cause most people to have high skin conductivity levels and to have a feeling of horror and revulsion, these patients will report in a cool cognitive way that the scenes are horrible and revolting, but they will not have any such feelings, nor will they have any measurable skin conductivity change. Their emotional detachment is remarkable, and might seem a feature, if it were not for the serious problems that such lack of emotionality actually seems to be a part of in day-to-day rational functioning, rendering these otherwise intelligent people severely handicapped. What these patients have is similar to what machines that coldly appraise emotions can have a level of emotion generation that involves appraisal, without any obvious level of bodily or somatic involvement.

It is not clear to what extent normal people can have emotions without having any associated bodily changes other than those of unfelt thought-patterns in the brain; consequently, the levels of emotion-generation described here may not typically exist in normal people without being accompanied by some of the mind-body linkages in the fourth component, described below. Nonetheless, multi-level generation of emotion is an important component because of its descriptive power for what is believed to happen in human emotion generation, and because some of these levels have already been implemented to a certain degree in machines. It is also relevant for certain neurologically atypical people such as high-functioning autistics who describe their ability to understand emotions as "like a computer – having to reason about what an emotion is" vs. understanding it intuitively.

The two levels just described – (1) quick and dirty sub-consciously generated emotions and (2) slightly slower, more reason-generated emotions, are not the only possibilities. Nor does my choice of these two examples impose a belief that "reasoning" has to be conscious. My point is instead that here are examples of emotions occurring via different levels of mechanisms. I expect that neuroscientists will find unique patterns of activation (and deactivation) across cortical and sub-cortical regions for each kind of emotion – joy, fear, frustration, anger, and so forth, and possible unique patterns for significant variations in levels of these. I would also expect we would build multiple levels of activation of emotion-generation mechanisms in machines, varying in resources used and varying in timing and in influence, in accord with the specific roles of each emotion. Some would be quick and perhaps less accurate, while some would be more carefully deliberated. Some would be at a level that could be consciously attended, or at least attended by some "higher" mechanisms, while some would occur without any such monitoring or awareness. Some of the mechanisms would be easy to modify over time, while others would be fairly hard-wired. Some of the emotion-generation mechanisms might be rule-based, and easy to reason about at least after the fact if not during, while others would be triggered by patterns of similarity that might not be easily explained. And many or even all of these mechanisms might be active at different levels contributing to background or mixed emotions, not just to a small set of discrete emotions. In summary, machines will have different combinations of mechanisms activating different emotions, a veritable orchestra for emotion generation.



Emotional Experience

What one can perceive of one's own emotional state:

I. Cognitive or semantic label
II. Physiological changes
III. Subjective feeling, intuition

Problem: consciousness

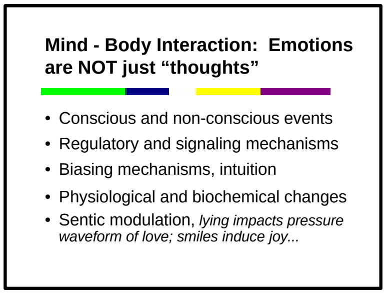
Emotional experience

We humans have the ability to perceive our personal emotional state and to experience a range of feelings, although many times we are not aware of or do not have the language to describe what we are feeling. Our feelings involve sensing of physiological and biochemical changes particular to our human bodies. (I include the brain and biochemical changes within it as part of the body). Even as machines acquire abilities to sense what their "bodies" are doing, the sensations remain different than those of human bodies, because the bodies are substantially different. In this *sense* machine feelings cannot duplicate human feelings. Nonetheless, machines need to be able to sense and monitor more of what is going on within and around their systems if they are to do a better job of regulating and adapting their own behavior. They will likely need mechanisms that perform the functions performed by what we call consciousness, if only to better evaluate what they are doing and learn from it.

A great distinction exists between our experience and what machines might have. The *quality* of conscious awareness of our feelings and intuition currently defies mechanistic description, much less implementation in machines. Several of my colleagues think that it is just a matter of time and computational power before machines will "evolve" consciousness, and one of them tells me he's figured out how to implement consciousness, but I see no scientific nuggets that support such belief. But I also have no proof that it cannot be done. It won't be long before we can implement numerous *functions* of consciousness such as awareness and monitoring of events in machines, but these functions should not be confused with the *experience of self* that we humans have. I do not yet see how we could computationally build even an approximation to the quality of emotional experience or experience of self that we have. Thus, I remain a skeptic on whether machines will ever attain consciousness in the same way we humans think of that concept. Consciousness, and life, for that matter, involves qualities that I do not yet see humans as capable of creating, outside of procreation. Perhaps someday we will have such creative abilities; nonetheless, I do not see them arising as a natural progression of past and present computational designs, not even with the advent of quantum computing.

If we can understand something, we can model it and build a computational model of it. Modeling is a form of imitation, not duplication. Thus, I use the term "imitate" instead of "duplicate" with respect to implementing this component in machines. In fact, we should probably be more careful about using the phrase "imitating some of the known mechanisms of human emotion in machines" to describe much of the current research concerned with "giving machines emotion." For brevity and readability the latter phrase is what I will continue to use, with hope that with this paper, we will begin to find some common understanding for what this shorter expression represents.

Mind-body interactions



Mind - Body Interaction: Emotions are NOT just “thoughts”

- Conscious and non-conscious events
- Regulatory and signaling mechanisms
- Biasing mechanisms, intuition
- Physiological and biochemical changes
- Sentic modulation, *lying impacts pressure waveform of love; smiles induce joy...*

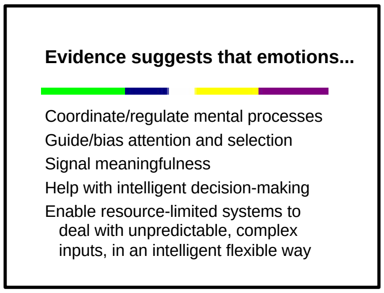
The fourth component is a broad category including many signaling and regulatory mechanisms that emotion seems to provide in linking cognitive and other bodily activities. Here, we find that emotions often involve changes in bodily systems outside the brain, as well as inside the brain. There is evidence, for example, that emotions inhibit and activate different regions of the brain, facilitating some kinds of cognitive activity while inhibiting others. Researchers have shown numerous effects of emotion and mood biases on creative problem solving, perception, memory retrieval, learning, judgment, and more. (See Picard (1997) for a description of several such findings.) Not only do human emotions influence brain information processing, but they also influence the information processing that goes on in the gastrointestinal and immune systems. (See Gershon (1998) for a description of information processing in the gut.)

Emotions modulate our muscular activity, shaping the space-time trajectories of even very simple movements, such as the way we press on a surface when angry vs. when joyful. I call the way in which emotions influence bodily activity *sentic modulation*, after Manfred Clynes’s work in sentics, where he first attempted to quantify and measure a spatio-temporal form of emotion. Clynes found that even simple finger pressure, applied to a nondescript firm surface, took on a characteristic pattern when people tried to express different emotions. Moreover, some of the emotions had implications for cognitive states such as lying or telling the truth. Subjects were asked to physically express either anger or love while lying or while telling the truth, and their physical expressions (finger pressure patterns, measured along two dimensions) were recorded and measured. When subjects were asked to express anger, the expressions were not significantly different during lying than when telling the truth. However, when subjects were asked to express love, the expressions differed significantly when lying vs. when telling the truth. In other words, their bodily emotional expression was selectively interfered with by the cognitive state of lying, given that it was not obviously interfered with in any other way.

I expect that this particular love-lying interaction is one of many that remain to be characterized. The interaction between emotions and other physical and cognitive states is rich and much work remains to be done to refine our understanding of which states inhibit and activate each other. As each interaction is functionally characterized in humans, so too might it be implemented in machines. Ultimately, if a machine is to duplicate human emotions, the level of duplication must include these many signaling, regulatory components of emotion, which weave interactive links among physical and mental states.

Consider building synthetic pain sensing and signaling mechanisms. Some machines will probably need an ability outside of their modifiable control to essentially *feel bad* at certain times, e.g., to sense a kind of highest-priority unpleasant attention-refocusing signal in a situation of dire self-preservation. (Where the word “self” is not intended to personify, but only to refer back to the machine.) This “feeling,” however it is constructed, would be of the same incessantly nagging, attention-provoking nature that pain provides in humans. When people lose their sense of pain, they allow severe damage to their body, often as the accumulation of subtle small damages that go unnoticed. Brand and Yancy (1997) describe attempts to build automatic pain systems for people, in one case a system that senses potentially damaging pressure patterns over time. The artificial pain sensors relay signs of pain to the patient via other negative attention-getting signals, such as an obnoxious sound in their ear. One of the ideas behind the artificial system is to provide the advantages of pain – calling attention to danger – without the disadvantages – the bad feelings. The inputs approximate those of real pain inputs, and the outputs are symbolically the same: irritating and attention-getting. Ironically, what people who use the artificial system do is either turn these annoying warnings off or ignore them, rationalizing that it can't be as bad as it sounds. Eventually the pain-impaired person gets seriously injured, although he or she doesn't really mind because it does not hurt. In short, the artificial pain system doesn't work; somehow it has to be "real" enough that you can't override or ignore it for long. Otherwise, injury accumulates, and the long-term prognosis is bad.

Whatever version of "pain" we give machines, if its goal is system-preservation, then it must be such that it is equivalent to not being able to be turned off, except under greater goals, or except by the machine's designer. This is not simply to say that pain-avoidance should always have the highest priority. Self-preservation goals may at some point be judged as less important than another goal, as suggested by Asimov's three laws of robotics, where human life is placed above robot "life", although this presumes that such assessments could be accurately made by the robot. Humans sometimes endure tremendous pain and loss of life for a greater goal. Similar tradeoffs in behavior are likely to be desirable in certain kinds of machines.



Evidence suggests that emotions...

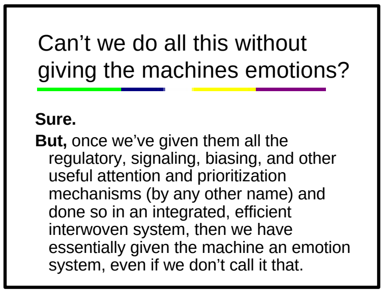
Coordinate/regulate mental processes

Guide/bias attention and selection

Signal meaningfulness

Help with intelligent decision-making

Enable resource-limited systems to deal with unpredictable, complex inputs, in an intelligent flexible way



Can't we do all this without giving the machines emotions?

Sure.

But, once we've given them all the regulatory, signaling, biasing, and other useful attention and prioritization mechanisms (by any other name) and done so in an integrated, efficient interwoven system, then we have essentially given the machine an emotion system, even if we don't call it that.

Before concluding this section, let me restate that it is important to keep in mind that all computers will not need all components of all emotions. Just like simple animal forms do not need more than a few primary emotion mechanisms, not all computers will need all emotional abilities, and some will not need any emotional abilities. Humans are not distinguished from animals just by a higher ability to reason, but also by greater affective abilities. I expect more sophisticated computers to need correspondingly sophisticated emotion functions.

Discussion

What is my view on what it means for a computer to have emotion? Before closing this discussion, we should keep in mind that we are still learning the answer to this question for living systems, and the machine is not even alive. That said, I have tried to briefly describe four components of emotion that are present in people, and to discuss how they might begin to be built into machines.

My claim, which opened this section, is that all four components of emotion occur in a healthy human. Each component in turn has many levels and nuances. If we acknowledge, say, N=60 such nuances, and implement them in a machine, then the machine can be said to have dozens of mechanisms of emotion that imitate, or possibly duplicate, those of the human emotional system. So, does the machine "have emotions" in the same sense that we do? There is a very basic problem with answering this: one could always argue that there are only N *known* human emotion mechanisms and more may become known; how many does the machine have to have before one will say that it has human-like emotions? If we require all of them to be identified and implemented, then one can always argue that machines aren't there yet, because we can never be assured that we have understood and imitated everything there is to know. Consequently, one could never confidently say that machines have emotions in the sense that we do. The alternative is to agree on some value of N that suffices as a form of "critical mass." But that is also ultimately unsatisfactory. Furthermore, some machines will have or may benefit from aspects of emotion-like mechanisms that humans don't have. Animals doubtless already have different mechanisms of emotion than humans, and we are not troubled by the thought of someone saying that they have emotions.

Ultimately, we face the fact that a precise equality between human and machine emotion mechanisms cannot be assured because we simply do not have complete lists of what there is to compare, nor do we know how incomplete our lists are.

Machines are still not living organisms, despite that we describe many living organisms as machines. It has become the custom to associate machine behavior and human behavior without really thinking about the differences any more. Despite the rhetoric, our man-made machines remain of a nature entirely different than living things. Does this mean they *cannot* have emotions? I think not, if we are clear that we are describing emotion as mechanisms with functional components like the four described here. Almost all of these have been implemented in machines at some level, and I can see a path toward implementing almost all of them. At the same time, it is prudent to acknowledge that one of the components, emotional experience, includes components of consciousness that have not yet been shown to be reducible to computational functions. Machines with all components but this one might be said to have emotion systems, but no real feelings.

As we make lists of functions and match them, let us not forget that the whole process of representing emotions as mechanisms and functions for implementation in machines is approximate. The process is inherently limited to that which we can observe, represent, and reproduce. It would be arrogant and presumptuous to not admit that our abilities in these areas are finite and small compared to all that is unknown, which may be infinite.

Remember that I began this presentation asking whether or not it was necessary to give machines emotions if all we are interested in is giving them the ability to recognize and respond appropriately to a user's emotion. Suppose we just want the computer to see it has annoyed someone, and to change its behavior so as not to do that again; why bother giving it emotions? Well, we still may not have to bother, if we can give it all the functions that deal with complex unpredictable inputs in an intelligent and flexible way, carefully managing the limited resources, dynamically shifting them to what is most important, judging importance and salience, juggling priorities and attention, signaling the useful biases and action-readiness potentials that might lead to intelligent decisions and action, and so forth. Each of these functions, and others that might someday be added to this list, may possibly someday be implemented by means other than "emotions." However, it is the case that these functions, in humans, all seem to involve the emotional system. We may find once we have implemented all of them, *and* integrated them in an efficient, flexible, and robust system, that we have essentially given the machine an emotion system, even if we don't call it that.

Machines already have some mechanisms that implement (in part) the functions implemented by the human emotional system. Computers are acquiring computational functions of emotion systems whether or not one uses the "e" word. But computers do not have human-like emotions in any rich or experiential natural sense. They may sense and label certain physical events as categories of "sensations," but they do not experience feelings like we do. They may have signals that perform many and possibly all of the functions performed by our feelings, but this does not establish equivalence between their emotion systems and ours. Computers may have mechanisms that imitate some of ours, but this is only in part, especially because our bodies differ and because so little is known about human emotions. It is science's methodology to try to reduce complex phenomena like emotions to a list of functional requirements, and it is the challenge of many in computer science to try to duplicate these in computers to different degrees, depending on the motivations of the research. But we must not be glib in presenting this challenge to the public, who thinks of emotion as the final frontier of what separates man from machine. When a scientist tells the public that a machine "has emotion" then the public concludes that not only could Deep Blue beat a grand master, but also Deep Blue could *feel* the joy of victory. The public is expecting that science will catch up with science fiction, that we will build HAL and other machines that have true feelings, and that emotions will consequently be reduced to a program available as a plug-in or free download if you click on the right ad. I think we do a disservice when we talk in such a way to the public, and that it is our role to clarify what aspects of emotion we are really implementing.

Emotions are not the system that separates man and machine; the distinction probably lies with a less popular concept -- the soul – an entity that currently remains ineffable, but is something more than a conscious living self. I don't have much to say about this, except that we should be clear to the public that giving a machine emotion does not imply giving it a soul. As I have described, the component of emotional experience is closely intertwined with a living self and I remain uncertain about the possibility of reducing this component to computational elements. However, I think that the other three components of emotion are largely capable of machine implementation.

If the day comes that scientists think that human emotion and its interactions with the mind and body are precisely described and understood, then it will not be many days afterward that such functions will be implemented in a machine, to the closest approximation as possible to the human system. In that time, most people would probably say that the machine has emotions, and few scientists will focus on what this means. Instead, we will focus on how machines, made in our image, can be guided toward good and away from evil, while remaining free to do what we designed them to do.

References

V. Braitenberg, *Vehicles: Experiments in Synthetic Psychology,* The MIT Press, Cambridge, MA, 1984.

P. W. Brand and P. Yancey, *The Gift of Pain*, Zondervan Publishing House, 1997.

M. Clynes, S. Jurisevic, and M. Rynn. Inherent cognitive substrates of specific emotions: Love is blocked by lying but not anger. *Perceptual and Motor Skills*, 70:195-206, 1990.

M. Clynes, M., Sentics: The Touch of the Emotions, Anchor Press/Doubleday, 1977.

A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain.* Gosset/Putnam Press, New York, NY 1994.

M. D. Gershon, *The Second Brain: The Scientific Basis of Gut Instinct and a Groundbreaking New Understanding of Nervous Disorders of the Stomach and Intestines*, HarperCollins, New York, 1998.

B. Johnstone, Japan's Friendly Robots, *Technology Review,* pp. 63-69, May/June, 1999.

J. E. LeDoux, *The Emotional Brain*, Simon & Schuster, New York, 1996.

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

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

R. W. Picard, Does HAL cry digital tears? Emotion and computers. In D. G. Stork, editor, *HAL's Legacy: 2001's Computer as Dream and Reality*, The MIT Press, Cambridge, MA 1997.

I. J. Roseman, A. A. Antoniou, and P. E. Jose. Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion*, 10(3): 241-277, 1996.

W. G. Walter, An Imitation of Life, *Scientific American*, pp. 42-45, May, 1950.