

Changing the *Subject*?

Does the turn from (artificial) intelligence to (artificial) emotions change the *Subject*?

Some ethical concerns

*

Naomi Sussmann
Haifa University

*

Much of the work currently done in the intersection between computing and emotions is in response to the experience accumulated over the past half century or so with artificial intelligence. It reflects deepening understanding of the field, its possibilities and potential but, no less, growing frustration with its limitations and the attempt to push forward, beyond them. But its connection to the more general intellectual and political background should not be overlooked either. Over the past half century or so, a growing realization of the limits of human reason, and of purely rational decision-making at that, encouraged a new appreciation of other, especially emotional and communicative aspects of human nature. So I think the best way to think of the present willingness of computer geeks to rethink the role of emotions in computing, is to see it in this larger context, where it is evidenced across disciplines today. It reflects a new understanding of the place and role of emotion in all aspects of human life, including functions considered essential to intelligence – such as learning, reasoning, decision-making and creativity.

*

Generally speaking, computer science is roughly divisible into two, distinct philosophies or world-views: computer-centered, and human-centered. The computer-centered approach to computing is that, which seeks to create "machines with general intellectual capabilities far outstripping those of humans."¹ It is primarily concerned with the scientific and technological progress required and implied by the effort to accomplish such super-intelligence. The human-centered approach, on the other hand, is that, which perceives computers instrumentally, as tools in the service of humankind. Its interest in the improvement and expansion of computers' capacities is tightly related to their potential contribution to this enterprise. Of course, this distinction is very schematic and so, never as clear cut in reality: it is better thought of as describing the two opposite ends of a single, more or less continuous sequence: computer science.

The emotions: a crossroads in computer science?

As it now seems, the intersection between emotions and computers is likely to become the sort of juncture, at which old coalitions collapse and new ones, form. Thus, it is not clear that computer-centered research should necessarily side with any and all applications of emotions to computers nor, conversely, that human-centered research will or should oppose any of them which do not, as yet, seem clearly advantageous to human interests. In fact, at present so much is still obscure, that the distinction itself, between a human-centered and a computer-centered approach, has

¹ Bostrom, Nick. "Ethical issues in advanced artificial intelligence," at: <http://www.nickbostrom.com/ethics/ai.html> (viewed: November 4th, 2006).

collapsed and all are equally blinded by the challenge.² Much of its resolution will be determined of course by new evidence gradually emerging from neuroscience,³ but most still depends on the theory of emotion by which any given research—indeed, research in neuroscience as well as in computer science—will be guided.

The absence of theory

But here's the crux: as yet, there's no such single, dominant theory nor, therefore, such guidance or obvious conceptual framework to refer to either.

Put most succinctly, the problem is this: Is emotion irreducibly, perhaps even constitutively subjective, in which case it requires a self or *subject* to be genuinely experienced by or validly ascribed to? Or is emotion, rather, an essentially behavioral or psychophysical phenomenon and thus, objective: conclusively and reliably detectable by reference to clear, measurable data? For if emotion is the former, then clearly, computers *cannot* be the subjects of emotion nor, therefore, be made genuinely intelligent – not, at any rate, unless and until they are made genuine *subjects*. But if emotion is the latter, then there's good chance not only that, in due course, computers could be endowed with emotions—or again, at least something very much *like* emotions (call them c-emotions)—but that they would, subsequently, have to be thought of and, more importantly, *treated* as potential social actors: as subjects—with all attendant implications of course.⁴

Changing the subject?

This, then, is the gist of the problem currently faced in affective computing: Are computers potential social actors?⁵ Is it plausible to suppose that they will ever be (made) capable of interacting with us in a way even remotely similar to that, in which we interact with other humans or social beings? That they will be properly conceived of, not as mere objects or things but as proper social beings: as subjects?

Note, that I do *not* ask whether any of this is good or desirable or advisable for us to pursue?⁶ Whether it should be good or beneficial, either on balance or absolutely, to create computers that have emotions? This is because I presume that the question of what is good or beneficial for us is, in this particular connection, irrelevant—which is not to say, of course, that it is unimportant. On the contrary: it is *all* important. But it is one of those cases in which 'ought' is bound to be pushed out of the way of 'can' and 'can,' to be taken to imply 'will.'

Yet, given that the prospect of rendering computers proper subjects *is* still very far into the future, let me step backwards now, to an earlier stage of the present

² See: Rosalind W. Picard, *Affective Computing* (Boston, Mass: MIT Press, 1997).

³ Damasio; Picard; etc.

⁴ See: Michael Muller. "Multiple Paradigms in Affective Computing" in: *Interacting with Computers*, 2004: August.

⁵ Michael Muller formulates this in surprisingly similar terms (*ibid.*)

⁶ Nor, indeed, do I ask "If we want computers to be genuinely intelligent, to adapt to us and to interact naturally with us." But this is not because I – as Picard, whose formulation this is – consider the answer's being in the affirmative so obvious, as to be uninteresting. Rather, it is because, as explained above, I consider it plainly (albeit somewhat tragically) irrelevant. There will always be *enough* people who will want to find out if it is really possible: want to actualize it, not necessarily thinking of whether they truly want its consequences, or what those consequences may be, if thought through carefully enough.

discussion: Conceding that there is no single, clearly dominant theory of emotion on offer at present and, further, that most computer scientists are (justly and understandably) reluctant to advance one of their own, is it a good idea to proceed with affective computing regardless? Let me put the question more bluntly: In the absence of *some* theory of emotion, how can affective computing proceed? On what basis, precisely? With reference to what guiding principles, towards *what* as its end?

Is emotion, behavior?

On the whole, current affective computing proceeds on the implicit assumption that, for all practical purposes, emotions are validly determinable by reference to behavioral, psychophysical variables. This is *not* to say that it assumes that emotions *are* reducible to those variables. But it is to say that it considers whatever remains from the relevant reduction to *not* be *essential* to a proper, meaningful identification and understanding of emotion. Thus, emotion *x* need not be identified as behavior *y*, but the occurrence of behavior *y* *is* taken to be sufficient to indicate the presence of emotion *x*.

Now, we have neither the space nor the time to seriously consider this thesis here. But in fact, this is for the better—indeed, it is as should be. Because part of the point I am trying to make here, and what I'd love to make you think of when this talk is over, is precisely this: That this *is* a serious, substantive thesis about the emotions; that it currently serves to guide an important part of the work done in computer science; and that *this*, at least, is *not* as should be. That is, no serious, substantive thesis about something as serious and substantive as emotion should serve to guide any serious, substantive research or practical work in any field unless and until it has undergone the most rigorous process of consideration, reasoning, and justification. This is not to say, of course, that research cannot proceed before there is an absolutely well grounded and established theory about the relevant topic. But it is to say that it should not proceed without any theory at all.

Or else?

Well,

From no-theory to new theory?

Having no theory to work with normally means one of two things: It can mean, first, that *no* work is undertaken *at all*—not, at least, until an strong enough theory presents itself and encourages new research. Alternatively, it can mean that work is undertaken regardless, which is to say, regardless of there being no grounds to proceed on, concretely. Of course, sometimes this is simply impossible, so no work or progress is accomplished anyway. But often enough, procession and progress are made—as they evidently now are, in affective computing—and "no grounds" has a good chance of quickly becoming or, more probably, *being made* into, the "new grounds." That is, research in the relevant field develops a sufficiently powerful interest in treating its effective subject-matter—in our case, behavior—as *if* it were its proper, theoretical subject-matter—emotion—all along, eventually managing to help us forget our initial interest in the latter: its distinctiveness and importance for us.

*

Ethical concerns:

Ethos and Mores distinguished

I move now to discuss some ethical concerns relating to this effective theoretical turn: the turn from emotions to behavior.

Before I do so, it should be useful, however, to clarify the exact sense of "ethical" referred to here: I use "ethical" here in its original, Greek, sense of *ethos*—as distinguished from its currently prevalent moral usage. Thus, I am not concerned with the *mores* advisable for maintaining "good practice and standards" in the intersection between emotions and computing: the particular restrictions, regulations, and norms that should or should not be imposed in this enterprise. Instead, I focus on the implications of the behavioral bias inherent to this intersection, for the human *ethos*—namely, for "the disposition, character, or fundamental values distinctive of humanness" as we know it": for "its characteristic spirit as manifested in its attitudes and aspirations."

Ethics precedes morality

To be sure, this is *not* because I doubt the importance of such social and moral policing in this, as any context. Rather, it is because I believe that, both logically and practically, ethics, in the sense just indicated, *must* antecede morality: For if we are not clear about who we are and who we want to be, how can we determine what may or may not be permitted in this or that context, technological or otherwise? If we do not examine and debate the sort of dispositions, character, or attitudes we think distinctive of humanness, how can we determine what should be preserved and encouraged and what, discouraged and guarded against instead? All these are ethical choices, not moral ones: so before we approach morality, we should clarify our ethics.

Further, in a more practical vein, issues of good practice and standards arising in the intersection between emotions and computing strike me as insufficiently different, or different in insufficiently interesting ways, from those arising in connection with other forms of computing or, alternatively, other forms of emotional interaction: thus, even what *is* unique in this context should be resolvable by some combination between those other two. Last, but by no means least, as has been repeatedly indicated above, my main concern here is with the gap opened by affective computing between the 'normal' usage of the term "emotion" and its specifically computer-scientific usage: thus, my ethical concerns derive not only from my anxiety about its almost casual reduction of emotions to behavior but also from its being done without being explicitly said or accounted for: indeed, this is the general ethical concern hovering over this talk until now—but now, let us move on, to more specific issues.

Computer recognition and response to emotions

I limit myself here to computers' and others artificial intelligence systems' capacity *only* to recognize and respond to emotions. By saying "only" I do not mean that this is minimal or trivial but rather, that I do not discuss the possibility that computers may themselves have or experience emotions. This is because, I admit outright, I am deeply committed to the view on which consciousness is of the very essence of emotions. So computer recognition and response to emotions seems to me to represent the limits of the problem. Indeed, if and when computers will have been given emotions of themselves, I think the distinction between them and humans as well as

the distinctiveness of humans as such, will have to be thoroughly considered and the nature of the entire dilemma, entirely altered anyway. So the discussion that follows is limited to the possibility that computers can "only" recognize and respond to the emotions of others, leaving aside also the entirely nontrivial question, whether it is possible to adequately respond to emotions without having or experiencing emotions oneself.

What does it mean for a computer to have the capacity to recognize and respond to emotions? What do such recognition and response mean and what, more specifically, may they serve for? Why should we be so keen on materializing them? Indeed, *should* we be so keen?

Possible applications

In her path-breaking book, *Affective Computing*, Rosalind Picard provides a long and fascinating list of uses to which affective computing could potentially be put: thus, she mentions affective mirrors; classroom barometers; listening to what you like; agents that know your preferences; or moods; sensitive toys; and so on and so forth. What is common to all these inventions is of course their innovativeness and the level of sophistication involved in their development. (Indeed: I wish I had half the imagination necessary to think of even half these things!). But what is also common to them all is that they do not merely infer emotion from behavior: they also respond in like terms – which is to say that they respond to emotion in behavioral terms.

Thus, imagine a system that detects when its user is sad or anxious. Of course, it doesn't know *why* she's sad or anxious, or *what* she's sad or anxious about: it knows only *that* she is so. In fact, it doesn't even know that she *is* sad: it knows only that she looks and feels sad: Her smile is dim; her complexion is pale; her pulse is low, her blood pressure is x and the heart beats at a pace of y. All these, it is programmed to know, indicate sadness, or anxiety, or distress. Now, assuming that it is not itself capable of having or experiencing emotion of its own, it can obviously not respond sensitively to its user's emotions: it cannot ask her what she's sad over, examine the reasons, causes, or the objects of her symptoms either with or without her; nor can it reason with her about them, lend a shoulder, give a hug or, least of all, just shut up and wait until its talked to (unless of course it is explicitly told so, or turned off. But I shall get to this immediately below). So it will, inevitably, *respond*. How *can* it respond? It may suggest that she do x by presenting x (or something associated with it) to her; or it may attempt to divert her attention from what she does; or play some music, or a movie that it know she likes, or makes her happy. And so on and so forth.

Ceci est qua? Emotion? Mood? Feeling? Behavior?

Now, let's disregard the possibility that the emotion the system infers from its user's behavior (pulse, smile, perspiration, heartbeat rate, etc.) is exactly the one she is in fact now experiencing, say sadness. Let's also overlook the fact that the system will, inevitably, *respond* once the relevant emotion is detected: that it cannot but respond. What will the system respond to, exactly? Is it emotion, or mood, or feeling? Are emotion, mood, and feeling gratuitously synonymous? Further, are they all trivially synonymous *and* reducible to the same set of psychophysical, behavioral indicators? How is sadness distinguished from sorrow? from distress? from longing? fatigue? boredom? How is happiness different from joy? joy from content? content from love? love from affection? How often have you found yourself searching for the right word

to describe your emotion? How often did you then come to recognize that you in fact had several emotions mixed up, rather than just one? How often have you then found the right word to express just the one bothering you? Did you mind the difficulty or were you indifferent to it? And the attempt itself? How did it strike you? What do you take it to mean? But then, again, how often did you feel blue when you were happy? Happy, when in fact depressed, or distressed, or anxious? How often do you think of your emotions, know them, talk about them, are aware of them? How often are you oblivious to them? Dismissive of them?

Can emotions die?

My point is that, by treating emotions as if they were moods, and moods as if they were feelings, and feelings as if they were emotions, and then, too, bundling them all up, together, under a single set of given psychophysical indicators, the enormous complexity and variety of our emotions, and other states of mind, is overlooked and effectively also, in the long run, dismissed. Consequently and inevitably, the complexity of our emotional world and life will be compromised: first, we will become impatient with the relevant distinctions; then, we will just lose the ability to recognize them. And then, eventually, we will just lose them, have them die out on us, so to speak. Now, you may retort, as I did initially, that emotional depth or complexity will not, cannot be so readily erased: that if they are, indeed, as central to our being or wellbeing as suggested above, then they will persist or we will just resist the systems that make us give up on them—either by improving them, or by discarding them altogether. This may be true of course: it may be that the enterprise of affective computing will not succeed and we should not be too surprised if the reason for this will, in the end, indeed be that people resent them or feel them to be insufficiently sensitive to their needs. But the truth will more probably be more complex than this: some emotions and mental capacities and complexities will survive and even thrive, some others will be dimmed, and still others will just languish and finally, die. Emotions, like most other non-physical things, require careful tending and care: when these are not provided for, they will just mutilate or die, perish.

Think of the famous Greek *eudaimonia*, of which we know, but which we cannot more than try to vaguely connect to, recognizing that we will never fully grasp, let alone experience; or think of the Roman experience of *glory*, apparently very powerful, but very different to what we now refer to by the same name; or think, alternatively, of the relatively new notion of *guilt*, which neither the Greeks and nor the Romans nor probably most other sufficiently pre- and non-Judeo-Christian cultures experienced in any meaningfully similar manner.

Losses and gains

And this, of course, is just a handful of examples: there are more, and more complex ones, too, all of which suggest that emotions can and do in fact die, and that the more complex ones will probably be lost for ever. Of course, there's something to be gained in return for most such losses: probably not everything and certainly not anything exactly commensurable, but losses and gains do tend to eventually even out or, alternatively, be somehow compensated for. So what I am *not* saying here, is that we are bound to lose without gaining anything or even, anything commensurable, in return. What I *am* saying, instead, is that what we are bound to lose may be precisely that, which we currently identify our uniqueness and very *ethos* with—that is, "The dispositions, character, and fundamental values distinctive of humanness."

Something new and wonderful may emerge out of all of this; indeed, something so new and wonderful perhaps, as we cannot even start envisaging yet, just as some may think of our own, modern culture, emerging as it did from the ruins of the Greek, and the Roman, and the pre- and proto-Judeo-Christian ones. But what is different here, I think, is that, unlike our predecessors, the new era we are speaking of here is going to be—indeed, will *have* to be—entirely of our making, so we can roughly envisage at least the kinds of *losses* that we are going to suffer by entering it and somehow or other attempt to limit and control them: adjust the systems we wish to create to what we respect and cherish in, but also simply identify ourselves with, rather than adjusting ourselves to what those systems, or we as their creators, may be capable of.