

humaine

D10d

**White paper on ethical issues in emotion-oriented
computing**

October 2007



Date: 1st October 2007

IST project contract no.	507422
Project title	HUMAINE Human-Machine Interaction Network on Emotions
Contractual date of delivery	<i>September 2007</i>
Actual date of delivery	<i>1st October 2007</i>
Deliverable number	D10d
Deliverable title	White paper on ethical issues in emotion-oriented computing
Type	Report
Number of pages	9
WP contributing to the deliverable	WP 10
Responsible for task	P.Goldie
Author(s)	P.Goldie, R.Cowie, S. Doring, I. Sneddon, B. Carstens, P. Petta
EC Project Officer	Philippe Gelin

Address of person responsible: **Peter.Goldie@manchester.ac.uk**

The future of the ethics of emotion-oriented technology

Introduction

The aim of this paper, which was presented at the HUMAINE Summer School in August 2007, is to consider the future of the ethics of emotion-oriented technology in the light of what we know today.

One way of thinking about emotion-oriented technology (from now on, we will use the acronym EOT) is that it is just another kind of technology, and that there is nothing special about it so far as ethics is concerned. On this view, we are just faced with the usual difficulties with technology concerning judgements under uncertainty about risk (Kahneman Slovic and Tversky 1982). We are certainly faced with these difficulties, and they should not be underestimated. But what we want to do in this paper is to focus on what is *special* about ethics and EOT. We believe it to be special in that it raises four fundamental issues.

We will now consider these four fundamental issues and how they manifest themselves in various ways. We should emphasise that it is not our concern here to consider the ethical implications connected with any particular existing research project, or with the risks of ethically sound technologies being misused or abused through getting into the wrong hands. These are matters best dealt with by an Ethics Committee such as the one we have in HUMAINE, and which were reported on by Ian Sneddon, Chair of the Ethics Committee, at the Summer School. Rather, our concern is to stand back from particular technologies, and indeed from HUMAINE itself, to look more widely at the kinds of long-term ethical considerations that should be borne in mind by anyone involved in carrying out research in EOT, or in funding and monitoring such research (although we should add that all of the issues discussed in this paper have been debated at length during the four years of HUMAINE). This paper is, one might say, a *tour d'horizon*.

1. Uncharted conceptual territory

EOT is distinctive because we lack a sound conceptual framework for understanding emotion itself, and this conceptual deficit makes it difficult to think clearly about risk. For example, there may be risks involved in mobile phone technology, and these risks may be hard to assess, but there is not

also a lack of a firm *conceptual* grasp of what is involved, of what mobile phones are. And much the same applies to nuclear technology. In contrast to the nuclear example, it should be emphasised that the lack of a sound conceptual framework for the emotions is not confined to the lay public, but also affects philosophers and scientists working on the emotions—and, of course, on consciousness more generally, often these days called the last frontier of science. There is no settled consensus on many conceptual issues concerning consciousness and emotions, and there is no sign that one will emerge in the near future.

That conceptual uncertainty manifests itself in our attitude towards emotions in EOT—towards machines that, in some way, engage with emotionality. We are, these days, quite untroubled by computing machines. We perceive no threat to our humanity from, for example, supercomputers that can compute over highly complex material, often at speeds and with accuracy that ordinary mortal cannot aspire to. We are, however, troubled about two aspects of technology, which many people do intuitively feel threatened by. The first is where the technology has what might broadly be called the capacity for creativity, and in particular artistic creativity. The second is where the technology has the capacity for emotion and emotionality. (We will not address the creativity question, although we believe that there are important emotional aspects to creativity which may explain some of our difficulties here too.)

Of course, no technology has been devised so far that is capable of *having* emotion, rather than merely simulating emotion. But one of the manifestations of the lack of a framework for comprehension referred to above is that we are even uncertain as to whether such a thing is so much as possible—whether application of the concept of emotion has a place outside the human being and other animals. For example, it is disputed whether something made of metal and carbon fibre could ever be capable of emotion, or whether it is necessary for emotionality that whatever is to have it must be composed of the same kind of stuff as we are composed of (Searle 1992).

And, perhaps most fundamentally, it is disputed whether, if something is to *have* an emotion (to be afraid for example), that thing must also *feel* that emotion—to have (or at least be capable of having) certain feelings that are characteristic of that emotion.

The contrast can be put in terms of the more general contrast between two kinds of consciousness: what the philosopher Ned Block has called *access consciousness* and *phenomenal consciousness* (1997). Roughly, access consciousness is the kind of consciousness involved in mere cognition—information storage and processing for example. So, for example, the capacity of something to recognise a threat and to respond with evasive behaviour has access consciousness.

And, still as part of access consciousness, a more complex organism might also be capable of recognising its own internal states, such as the state which represents *that* it is threatened and *that* a certain kind of evasive response is called for. Phenomenal consciousness, in contrast, is what is involved when *there is something that it is like* for the organism—in this case, where there is something that it is like to feel fear (Nagel 1974). There is something that it is like to be a human, a dog, or a cow—they all have phenomenal consciousness, and they all can feel fear—but there is nothing it is like to be a stone, or a computer.

We do not know whether there could there ever be something that it is like to be a robot—could a robot ever feel fear? However, as science fiction literature and film attest, people can easily be disturbed by the idea of non-animal things that are capable of emotional feelings: consider, for example, the Nexus-6 replicants in *Blade Runner* who are programmed with a fail-safe device to cease functioning after four years in case they start to develop empathy (Goldie forthcoming); and Hal in *2001: A Space Odyssey*, who seems to be motivated emotionally, by revenge or envy perhaps, and who seems to suffer as his systems are shut down.

We cannot emphasise enough that these are thought experiments: the real work in EOT is very far from all this. Even to say that it is tomorrow's problem would be misleading, as we have no idea whether it is even conceptually possible that such things might exist outside science fiction and philosophers' thought experiments. However, because these issues go the heart of our humanity, people tend to have a nervousness about the more practical and feasible aspects of EOT that, perhaps unrealised by them, is coloured by the wider conceptual issues we have been discussing in this section. Any research project in EOT must be properly sensitive to these issues. Science fictional characters should not simply be dismissed as mere science fiction: science fiction they are, but, in respect of their emotional resonance to current research and technology, they are not *mere* science fiction.

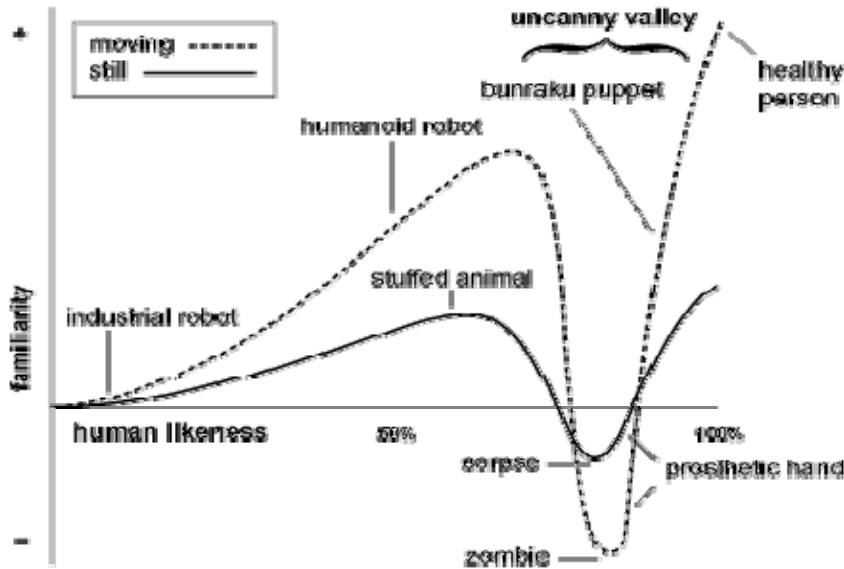
So the capacity for emotion is in some sense special for us humans. Quite what the sense is in which it is special is very hard to get a grip on, in large part because of the lack of a framework for comprehension just referred to. But even if we don't know quite what emotions are, nevertheless, we know that they are special. Emotions (and consciousness more generally) may be the last frontier for science, but they are also the last bastion of humanity.

2. The last bastion of humanity

We now put science fiction aside to consider people's confused attitudes towards *simulation* of humanity and emotionality in EOT. In other words, our concern now is not with EOTs that *have* emotion, but with EOTs that *give the impression of emotionality*. This is possible in various manifestations in robots, in avatars, and in ECAs, through speech, appearance, behaviour and in other ways.

First, we want to mention an important and often neglected point. In human-machine interaction, we are, and have long been, very familiar with the idea that we interact with machines that simulate emotion, and that are involved in bringing about emotional responses in the human user. Consider, for example, how emotionally-charged is our reaction to the pre-recorded telephone message from the airline company, telling us they are 'sorry' to keep us waiting, and that our call is 'valuable' to them. We often have strong negative emotions in response—frustration, anger, feelings of inadequacy—at this blatant pretence of caring from a 'system' in which we are at best a number on a screen. This kind of interaction with technologies seems to be a near inevitability in our lives these days. The neglected point, then, is that, given this state of affairs, it is an excellent thing that there are people working in EOT aiming to develop better systems that will reduce or even eliminate such negative emotions in our interactions with technology. The following remarks should be understood in that overall context.

In the particular case of robots, it has been argued that our emotional responses take a curious shape, in what is called *the uncanny valley*, a term coined by Mori (1970), and now much discussed in robotics and computer science. The essential idea, captured in the following diagram, is that our emotional attitude towards robots changes as they become more and more similar to human beings (in behaviour, in facial and verbal expression, and so on). The claim is that, as the diagram shows, the relation is not a linear one. We are more comfortable with a humanoid robot than an industrial robot, but when the artefact becomes close to being like a healthy human but is still clearly *not* human, our feelings of comfort and familiarity decline: we are in the uncanny valley.



(‘The Uncanny Valley’, Masahiro Mori, *Energy*, 7(4), 1970 pp. 33-35; adapted by Karl F. MacDorman and Hiroshi Ishiguro, *Interaction Studies* 7:3 (2006), 297–337.)

It is not clear that it is accurate to talk about an uncanny valley. The most systematic study of the topic suggests that it is not (Hanson et al, 2005). Creating progressively closer approximations to humanity *may* create artefacts that are ‘bukimi’ in Mori’s sense—weird, ominous, eerie—but it is not clear that it *usually* does so. When people react in that way to particular artefacts, it may be because they give rise to disgust; because they deviate from the norms of physical beauty; because they frustrate our (largely unconscious) expectations; because they give rise to fear of death (MacDorman and Ishiguro 2006). When these emotional responses occur, they are generally not of the kind that can be seen as rational, in the way that, for example, fear of a savage dog would be rational. They are, rather, more visceral, more primitive. Note, though, that people’s readiness to believe that there is an uncanny valley, on only the slightest evidence, suggests that visceral responses also penetrate what people believe is rational consideration of the issues. It seems likely that these responses are, at least in part, yet another expression of our confusions about EOT. They run through all discussions of the ethics of the issues.

There is a debate in robotics as to whether it should work towards developing a robot that we might, in our interactions with it, mistake for a human being, or at least treat in way that reflects confusion on our part about how it *should* be treated (perhaps as responsible autonomous moral agents). Recent research has been carried out to show that such confusion arises to some extent even in relation to avatars and robots that are *known* not to be human (Reeves and Nass 1996;

Slater et al 2006; Rosalia et al 2005; Bartneck et al 2006). However, the idea of a robot that a human being might genuinely confuse with another human being is, as yet, too far into the realm of science fiction to impinge on practical ethics at the present time.

3. The reflexivity of ethics and EOTs

Our ethical responses to EOTs are themselves emotional. The old idea that ethical judgements are cool and dispassionate has now been replaced, in philosophy and in psychology, with a picture where emotions are central to our ethical intuitions and judgements (Haidt 2001, 2007). So there is a tricky reflexive aspect to ethics and EOT: the tool that we are using in our ethical deliberations is the very tool that is under examination in those deliberations.

An immediate response to the question of what ethical stance we should take towards technology that merely simulates emotionality might be a dismissive one: such machines are of mere instrumental value, and should be treated no differently to the way we treat a can opener or a laptop computer. The idea that such machines could have rights, or that we could have duties towards them, might accordingly be thought insupportable, or even absurd. This may well be the correct reaction, at least so far as concerns rights and duties. But still, there might be good reasons to treat such machines as non-instrumentally valuable (Goldie unpublished; Pelachaud 2006). Our feelings towards, and the way we treat, EOTs, is expressive of our personality, and personality traits (of this kind) are largely a matter of habit. So there is a risk that we can become habituated in treating EOTs badly, and from this (especially bearing in mind the uncanny valley) there is a further risk that we will start to treat certain humans like this too, merely as means. This idea too is familiar from literature and film, and is often associated with a dystopia (Fritz Lang's *Metropolis* for example). On this view, then, we should cultivate our personality to make sure this doesn't happen, that we don't slide down the slippery slope to treating human beings in this way too.

We turn now to another way in which reflexivity of ethics and EOT is manifested. EOTs are capable, to an increasing degree, of using emotions to persuade users. They can 'use' emotions in two senses. The first sense is the one that is familiar to us through our encounter with TV and other advertisements: the way in which they can appeal to our emotional sensitivities to persuade us to act in certain ways—to buy a product, or to take a holiday somewhere. The second is less familiar: the way in which EOTs could simulate emotionality in themselves in order to generate emotional responses in the user, and thus persuade us to act in certain ways (as discussed by Marco Guerini for HUMAINE). As evidenced by the use of Tagamochi toys with children, this can be highly

effective, possibly in deleterious ways. A number of familiar ethical issues have application here. Let us mention just two. Issues arise concerning whether the end can justify the means. For example, if an EOT is more likely than a doctor to get true answers from patients to a medical questionnaire, would the end (better health for the patient) be justified by the means (rhetorical persuasion, perhaps subliminal, by the EOT)? And, secondly, issues arise concerning whether rhetorical emotional persuasive devices in EOTs undermine the autonomy of the user. For example, would a patient using a persuasive EOT justifiably consider his autonomy to have been undermined if she is not properly informed of the procedures (which might in itself eliminate their usefulness)? Again, as we saw earlier, perception of risk with regard to such issues itself involves emotion (Slovic 2007)—another aspect of reflexivity.

4. EOT, ethics and the law

What is legal and what is ethical are not, of course, co-extensive. But in many cultures and in many circumstances, the law will often embody our intuitions about what is right and wrong. There is in Western Europe little legislation that is specifically aimed at EOT, at least as far as we are aware (legislation on polygraphs or lie detectors being perhaps an exception). So any cases would have to be decided on existing legislation and case law, interpreted and applied as thought appropriate—for example, in employment legislation, human rights, and privacy laws. Difficulties may well arise here, in part because of the uncharted conceptual waters which we have already discussed. For example, if there is an EOT which has the capacity to recognize someone's emotion from their facial expression in public, is it an invasion of privacy to record emotions in this way?

This example brings to light a more general issue concerning EOT. Many EOT applications, and this is an example of one, are what are called SIIFs: semi-intelligent information filters. It is semi-intelligent in that it records more than just raw data—bodily posture and movements, facial expressions, eye saccades and so on: it also *interprets* this data in a meaningful way: for example, recording that the person is upset or feeling aggressive. SIIFs can be, and often are, enormously useful; for example, in-car technology can be used to determine whether a driver is safe to drive, and to advise him or her accordingly. As a filter, a SIIF will, as in this example, characteristically have the power to transmit data for further action. And this data can be used for a high-impact judgement about the person being observed. For example, a SIIF of the kind envisaged might be useful in monitoring employees in a call centre, or for use by anti-terrorist police in monitoring people in crowded public areas, such as shopping malls or railway stations.

As these examples illustrate, the potential for good of SIIFs is substantial, although there are ethical issues that arise, some of which are analogous to those that have arisen with polygraphs: for example, it is important to avoid exaggerated claims of accuracy, and to ensure that questions of admissibility as evidence and invasion of privacy are properly addressed.

Those working in EOT, as researchers, or as funding bodies and monitors, are well aware of these issues. And it is right that they should be, because there is always a risk that EOT will come into the public gaze at times when the technology has been abused, or at least when such abuse is being alleged, and this could affect the public attitude towards a research programme that aims to do good.

5. Conclusion

We are in uncharted conceptual waters with EOT, and emotion is seen as the last bastion of humanity. Accordingly, emotions run high about the ethics of EOT (the reflexivity point), and it is essential for us all to be sensitive to this. But it must not be forgotten that EOT is an enormous force for good, for example in reducing or eliminating the negative emotions that we so often feel in our interactions with complex technologies, from computers to car navigation systems to online questionnaires and booking services. Those working in the field should try at every opportunity to proselytise the benefits of EOT to humanity: to make our lives easier and better. And at the same time, it is essential that there are in place adequate systems for ethical governance of the kind that we have in HUMAINE, able to draw on real depth and breadth of expertise in science, in emotion theory, and in ethics.

Bibliography

- Bartneck, C. 2006. 'To Kill a Robot'. *Proceedings of the Workshop on Misuse and Abuse of Interactive Technologies in cooperation with the Conference on Human Factors in Computing Systems (CHI2006)*, Montreal.
- Block, Ned. 1997. 'On a Confusion about a Function of Consciousness', in *The Nature of Consciousness*, eds. Block. N., Flanagan. O., and Guzeldere. G., Harvard, Mass.: MIT Press.
- Goldie, Peter. Forthcoming. 'What is it like to be a Nexus-6 replicant?', for a collection on the film *Blade Runner* edited by Amy Coplan for Routledge.
- Haidt, Jonathan. 2001. 'The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgement'. *Psychological Review* 108, 814-34.
- Haidt, Jonathan. 2007. 'The New Synthesis in Moral Psychology'. *Science* 5827, 998-1002.

- Hanson, D., Olney, A., Prilliman S., Mathews, E., Zielke, M., Hammons, D., Fernandez, R., and Stephanou, H., (2005) “Upending the uncanny valley”, in Proceedings of the Twentieth National Conference on Artificial Intelligence, Menlo Park, CA: AAAI Press, pp. 1728–1729.
- Kahneman, Daniel, Paul Slovic, Amos Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
- MacDorman, Karl F., and Ishiguro, H. 2006. *Interaction Studies* 7, 297–337.
- Mori, Masahiro. 1970. ‘The Uncanny Valley’. *Energy*, 7, 33-35.
- Nagel, Thomas, 1974. ‘What is it like to be a bat?’. *The Philosophical Review*. 83, 435-450.
- Nass, C, and Reeves B. 1996. *The Media Equation: How People Treat Computers*. Cambridge: Cambridge University Press.
- Nisbett, R., Ross, D. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs: Prentice-Hall.
- Power, Samantha. 2003. *A Problem from Hell: America and the Age of Genocide*”. New York: Harper Perennial.
- Rosalia, C., Menges, R., Deckers, I., & Bartneck, C. 2005. Cruelty towards robots. *Robot Workshop - Designing Robot Applications for Everyday Use*, Göteborg.
- Searle, John. 1992. *The Rediscovery of the Mind*, Harvard, Mass.: MIT Press.
- Slater M, Antley A, Davison A, Swapp D, Guger C, et al. 2006. A Virtual Reprise of the Stanley Milgram Obedience Experiments. *PLoS ONE* 1(1).
- Slovic, Paul. 2007. “‘If I look at the mass I will never act’: Psychic numbing and genocide”. *Judgment and Decision-making*, 2, 1-17.