The Ethical Distinctiveness of Emotion-Oriented Technology: Four Long-Term Issues

Peter Goldie, Sabine Döring, and Roddy Cowie

Abstract In this chapter we consider certain long-term ethical issues which are peculiar or special to emotion-oriented technology and which make the topic particularly charged ethically, both for the lay public and for those working in the area. We identify four such issues. First, it is far from clear whether technologies made by humans are conceivably capable of emotionality and, more generally, of phenomenal consciousness. Second, where we are dealing with a technology that simulates emotionality, we have responses that are often far from cool and rational. Third, as discussion of the first two issues will have illustrated, our ethical responses to emotion-oriented technology are often emotionally charged, lending a peculiar reflexivity to our ethical deliberations. This leads to a discussion of the kind of value that such technologies have, and of how they should be ethically treated by humans. Fourth, emotion-oriented technology impinges on many matters of law (such as laws of privacy); we discuss in particular the importance of technology that is used to filter raw emotional data on people for further use.

1 Introduction

The aim of this chapter is to consider what is special about the ethics of emotion-oriented technology.

One way of thinking about emotion-oriented technology (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 et al., 1982). We are certainly faced with these difficulties, and they should not be underestimated. But there is a good case for believing that EOT raises issues that are not covered in every text that deals with ethics and technology.

P. Goldie (✉)
Department of Philosophy, University of Manchester, Manchester, UK
e-mail: peter.goldie@manchester.ac.uk

P. Petta et al. (eds.), *Emotion-Oriented Systems*, Cognitive Technologies,
DOI 10.1007/978-3-642-15184-2_39,

If there are special ethical issues to be faced, then it makes sense for the community to understand them in some depth. This chapter is directed towards people who take that view seriously. Certainly for most computer scientists (or engineers or even psychologists) working in the area, it will be possible to cope most of the time without thinking more deeply about ethics than the minimum practical requirement ensuring proper scrutiny by an ethical committee. But there are also times when the community is challenged on ethical grounds, and then it needs people who can engage with the non-routine issues that this chapter raises.

We propose that what makes the area distinctive rests on four fundamental issues. The chapter will consider these four fundamental issues and various ways in which they manifest themselves. 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 that can and should be dealt with by an ethics committee.

The issues that this chapter deals with are the obstinate, vaguely defined concerns that seem troubling to the general public, who are not generally deeply informed about the technical details of research that ethics committees deal with; and to the individuals and bodies involved in funding and monitoring such research. It is, one might say, a *tour d'horizon*.

2 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 mortals 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.)

One of the manifestations of the lack of a framework for comprehension referred to above is that we are uncertain whether it is at all imaginable for technology to have the capacity for emotion. It is quite possible to argue that the concept of emotion has no 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 (1995). 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). Clearly 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 it seems obvious to most people that there is nothing it is like to be a stone, or a computer.

We do not know whether there could 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 4 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.

The point of these examples is to underline the slightness of the conceptual framework that human beings in general bring to understanding emotion itself. The examples are based on 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. But the experiments are about issues that go the heart of our humanity, and they expose uncertainty about those issues. It is no wonder, then, that people tend to have a nervousness about the more practical and feasible aspects of EOT. They may not know the reasons, but it

is reasonable to be nervous given the wider conceptual uncertainties that 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. The science fiction is a reflection of the fact that 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 we do not need to know exactly (or even roughly) what emotions are, to know that they are special. We have said that emotions (and consciousness more generally) are the last frontier for science. But unlike other frontiers of science – the unification of relativity and quantum theory, and so on – they are also personally significant: the last bastion of humanity.

3 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. Our concern shifts from the elusive notion of EOTs that have emotion, to the likely future reality of EOTs that give the impression of emotionality. This is likely to be possible in various manifestations in robots, in avatars and in embodied conversational agents (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 Fig. 1, 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

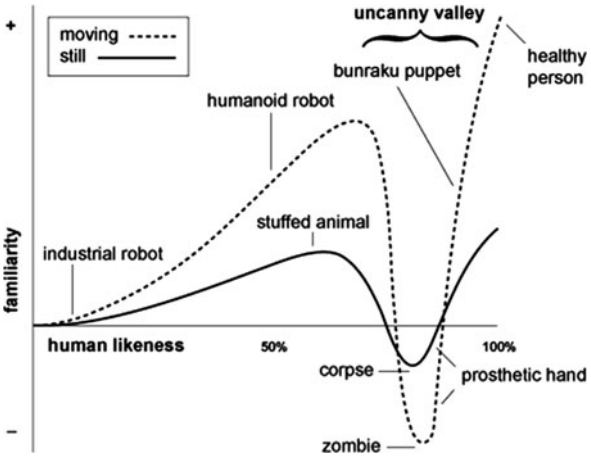

Fig. 1 'The uncanny valley' (Mori, 1970); adapted by MacDorman and Ishiguro (2006)

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 concept of an uncanny valley has an immediate appeal, but it is not clear that it is accurate. 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. Not everything that comes close to human appearance is like a corpse, or a prosthetic hand, or a zombie. These very particular human-like things may signal potholes that a sensible technologist can and should steer round rather than a long and unavoidable valley.

Even if that is so, the potholes need to be understood. 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. Technology is both ethically and practically bound to deal carefully with these responses and to make sure that they are not roused inadvertently.

Furthermore, 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 indicates that such confusion arises to some extent even in relation to avatars and robots that are known not to be human (Reeves and Nash, 1996; Slater et al., 2006; Rosalia et al., 2005; Bartneck et al., 2006). It may therefore be less difficult than one might assume to build a robot that a human might genuinely confuse with another human being. It is, as yet, far into the realm of science fiction, and so it does not impinge on practical ethics at the present time. But like uncertainty about emotion itself, it is an obstinately disturbing echo that disrupts attempts to hold cool, rational discussion about EOT.

4 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 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; Bartneck et al., 2008). 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 does not happen; that we do not slide down the slippery slope to treating human beings in this way too.

There is 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, second, 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 he 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.

5 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 recognise someone's emotion from their facial expression in public, is it an invasion of privacy to record emotions in this way?

This example illustrates an important general issue concerning EOT. There are enormous practical complexities in deciding what is consistent with the law as it currently stands, and in seeing how the law could be adapted to allow applications that are benign, and rule out those that are not. Behind those complexities lie ethical questions about real devices rather than imaginary futures, which hinge as much on use as the device itself.

The issue is well illustrated in a class of applications that have been called SIIFs: semi-intelligent information filters. The hypothetical emotion recogniser mentioned above is a case in point. 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 they can be used to determine whether a driver is safe to drive, and to advise him accordingly. As a filter, a SIIF will, as in this example, characteristically have the power to transmit data for further action. The same data, though, 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. In some of those situations, there is a real risk that a false positive could affect a person's life, or end it.

As these examples illustrate, even SIIFs with clear potential to do good raise complex issues on the borderline between ethics and the law, some of which are analogous to those that have arisen with polygraphs: for example, how to ensure proper understanding of a device's 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 potential for abuse of EOT. There is always a risk that it will come into the public gaze at times when abuse has happened, or at least when it is being alleged; and this could affect the public attitude towards a research programme that aims to do good.

6 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. It is part of the task for people working in the field to proselytise the benefits of EOT to humanity: to explain how it can 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 were introduced in HUMAINE, able to draw on real depth and breadth of expertise in science, in emotion theory and in ethics.[1]

References

Bartneck C, et al. (2006) 'To Kill a Robot'. In: Proceedings of the workshop on misuse and abuse of interactive technologies in cooperation with the conference on Human Factors in Computing Systems (CHI2006), Montreal, QC

Bartneck C, Brahnam S, De Angeli A, Pelachaud C (2008) Misuse and abuse of interactive technologies. Interact Stud – Social Behav Commun Biolog Artif Syst 9(3):397–401

Block N (1995) On a confusion about the function of consciousness. Behav Brain Sci 18:227–287

Goldie P Forthcoming. What is it like to be a Nexus-6 replicant? In: Coplan A (ed) Blade runner. Routledge, London

[1] For those who are interested in an outline to some of the current philosophical issues in emotions, a good place to start is Ronald de Sousa's entry in the online Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/entries/emotion (accessed 17 May, 2010)

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgement. Psychol Rev 108:814–834
Haidt J (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, Stephanou H (2005) Upending the uncanny valley. In Proceedings of the 20th national conference on Artificial Intelligence, AAAI Press, Menlo Park, CA, pp 1728–1729
Kahneman D, Slovic P, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge, MA
MacDorman KF, Ishiguro H (2006) The uncanny advantage of using androids in cognitive science research. Interact Stud 7(3):297–337
Mori M (1970) The uncanny valley. Energy 7(4):33–35
Nagel T (1974) What is it like to be a bat? Philos Rev 83:435–450
Reeves B, Nash C (1996) The media equation: how people treat computers. Cambridge University Press, Cambridge, MA
Rosalia C, Menges R, Deckers I, Bartneck C (2005) Cruelty towards robots. In: Robot workshop – Designing robot applications for everyday use, Göteborg
Searle J (1992) The rediscovery of the mind. MIT Press, Harvard, MA
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):e39
Slovic P (2007) 'If 'I look at the mass I will never act': psychic numbing and genocide. Judgm Decis Mak 2:1–17