Ethical issues in affective computing

Roddy Cowie

Abstract

Affective computing is bound up with ethics at multiple levels, from codes governing studies with human participants to debates about the proper relationship between ethical and emotional systems within an agent. Behind the debates lie ethical principles which are powerful, but divergent. In some areas (e.g. data protection and research with human participants) explicit codes give them legal force. Elsewhere, they give rise to characteristic imperatives: to increase net positive affect, to avoid deception, to respect autonomy, to ensure that systems' competence is understood, and to provide morally acceptable portraits of people. There are also widely discussed concerns with less clear connections either to classical ethics or to the real abilities of the technology, but they still need to be addressed.

Keywords
Emotion, affective computing, ethics, ethical codes, moral philosophy

1. Introduction

People who work in affective computing tend to have trained in disciplines allied to engineering and mathematics. Training in those areas is unlikely to have included courses on ethics. As a result, it can come as a shock to discover that ethical issues are very much part of the discipline that they have come into; and at several levels, not just one. At the most specific, when they collect data from human participants, they need to arrange an acceptable form of ethical approval. At the most general, they may find themselves pressed to answer high-profile claims that the whole enterprise of affective computing is ethically tainted. Particular applications, from artificial companions for the elderly to sex robots, pose various individual difficulties.

The aim of this chapter is to give people a grounding that lets them engage with that range of challenges in a rational way. There is no simple way to do that. The approach that the chapter takes is to provide some general conceptual background to begin with, and then to look at topics with a specific bearing on affective computing. Broadly speaking, it begins with the topics where the ethical concerns are most clearly-defined, and works towards those where the issues are hardest to articulate precisely. Particularly in the last group, the arguments are often about what people imagine a computer with emotions or emotion-related skills might be like, rather than anything that can be built, or realistically envisaged. That does not mean that those arguments can be ignored. In fact, they may be the most important for the future of the discipline.

People working in affective computing may well feel that the coverage overemphasises particular parts of the discipline. That is essentially because a handbook chapter has to reflect the balance of the relevant literature. The literature says more about production rather than perception – at least partly because ethical theory has traditionally focussed on evaluating actions, and so its obvious application is to the actions that a system might take. However, the chapter does what it can to engage with less traditional issues that are important to the discipline, such as the way it portrays human beings.

2. Formal and informal foundations of ethics

The discipline that discusses the foundations of ethics has traditionally been called moral philosophy. There is no universal agreement on the use of the terms 'ethics' and 'morality', but it is reasonable to adopt the convention that morality includes ethics. In that sense, calling a judgement ethical implies that it is moral, but it also implies that it is grounded in reason, rather than just a gut feeling for or against.

Nevertheless, the gut matters in ethics, and it is important for affective computing that it does. What we would call emotion was identified as the root of moral judgement by thinkers with a huge influence on the modern era, notably David Hume (1740) and Adam Smith (1759). For them, doing wrong is ultimately about producing situations that are unacceptable to our ‘moral sentiments’. Later thinkers, notably Bentham and John Stuart Mill, developed a very well known form of the idea, the ‘utilitarian’ principle that our fundamental moral duty is to bring the greatest happiness to the greatest number (Driver, 2012). Whether or not one agrees with the arguments, they reflect a deep-seated intuition that emotion is at the core of what makes us moral beings. For that reason alone, nobody should be surprised that ethics cannot be kept out of affective computing.

Emotion-based theories contrast with two classical alternatives. One is associated with Kant (Wood, 1999). He argued that the bedrock of morality was exercising free will in accordance with intellectual principles rather than feelings. Specifically, we should act in accordance with principles that we could will all rational agents to follow. For those who accept that idea, affective computing has profound moral significance, because it raises the prospect of creating things that mimic human free will, or impinge on it.

The last major alternative, famously advocated by Hobbes (1651), is that moral codes are contracts established for the purpose of maintaining a society that satisfies our basic desires. For Hobbes himself, the contract can and should be imposed by a strong authority. For others, it should emerge from shared values (Scanlon, 1998). People who set rules are quite likely to assume that Hobbes was broadly right.

These approaches are particularly important for affective computing, but many others are comparably important from a purely philosophical perspective. Examples include the revival of virtue ethics (McIntyre 1985), the argument that moral claims are simply errors (Mackie 1977), sophisticated attempts to reconcile the major positions (Parfitt, 2011), and much more. The details are fascinating, but probably not mandatory reading for people in affective computing.

Attempts to ground ethics in fundamental principles have two features which are important and troublesome for affective computing. First, many people regard one or more of them as self-evident, and overwhelmingly important. For example, the concept of autonomous beings with free will seems self-evident, intellectually profound, and effectively sacrosanct to a great many people. Second, they lead to disputes that no amount of rational argument will resolve, because what one party takes as a self-evident starting point seems opaque and counter-intuitive to the other. Part of a sophisticated ethical stance is understanding that those difficulties are bound to arise when we try to argue from first principles.

Because that problem is well known, philosophers who work in practical ethics have developed systems that are closer to common sense, and more likely to promote consensus. The approach was pioneered by W.D. Ross (1939). He aimed to identify a few basic principles that speak for or against a course of action. The best known list consisted of *fidelity* (a duty to keep our promises); *reparation* (a duty to right a wrong we have done); *gratitude* (a duty to benefit those from whom we have accepted benefits); *non-maleficence* (a duty not to harm others); and a duty to maximise the aggregate of good. Perhaps the most eminent philosopher to have written about ethics and affective computing, Peter Goldie (Döring, et al, 2011), directed the field to Principalism, a related approach proposed by Beauchamp and Childress (2001). Their list includes Ross's last two items, non-maleficence and benificence. The other items have a more Kantian flavour: they are autonomy (i.e. to promote rather than restrict people's ability to exercise free will); and equity (that is, not to treat people differently for no good reason).

An older, and even more compact summary is the ‘Golden Rule, “do as you would be done by”. There are well-known problems with the Golden Rule (Blackburn, 2001) – it is famously not a good prescription for judges or masochists, and perhaps people who love programming should be added to the list. But there is a widespread sense that it is essentially sound, and it has been explicitly applied in fields allied to affective computing (Berdichevsky & Neuenschwander 1999).

On the other hand, various documents which are called codes of ethics are essentially Hobbesian statements of what people in authority require of people under their authority . They often reflect views with very broad support, but sometimes they take highly contentious positions. For example, the British Engineering and

Physical Research Council drew up ‘ethical rules for robotics’[1]. These seem to stipulate that however closely robots’ intelligence and emotions might approximate ours, the fact that they are manufactured requires them to be regarded as tools, and perhaps to display bar codes declaring their status. To many, that seems quite the opposite of ethical.

The key point here is that following the dictates of one’s own conscience and reason will not necessarily mean staying on the right side of the ethical codes that authorities establish. The two can and do diverge. Usually conforming is harmless. Deciding what to do when it seems genuinely wrong to conform is a notoriously difficult problem.

Public opinion raises similar issues. Quite possibly the biggest threat to affective computing is that the public may come to feel that it is ethically unacceptable. Nobody should doubt the importance of finding ways to counter the unease, and certainly to avoid heightening it. The concerns to be countered, though, often seem to involve moral principles that are not traditionally central to moral philosophy.

The most obvious of the principles is that certain kinds of unnaturalness are bad. Since civilisation is based on unnaturalness, the concern is obviously to do with specific kinds of unnaturalness rather than unnaturalness in general. It may be to do with the way we think about things that are living, or that behave at some level as if they were living: that would fit a line of argument developed by Foot (2001), which suggests that we have particularly strong intuitions about the way living things should or should not be. Be that as it may, there is clearly a widespread feeling that a computer that seems to have emotions is unnatural in a morally disturbing way.

A second principle is that there are parts of existence where it is morally wrong to venture. However people may rationalise it, the feeling is clearly akin to a religious one: the ground is sacrosanct, and treading there is sacrilege. We may or may not agree that emotion is an inner sanctum of humanity, but clearly, a feeling of that kind comes into play when people judge how ethical or unethical the enterprise of affective computing is. Finding an effective answer to that reaction depends on understanding it.

A third principle, less deep, but still to be reckoned with, is that certain kinds of frivolity are bad. Technology in particular should be concerned with real problems, like storing more information, or solving problems faster. Making the user feel better is not a fitting use for technical skills and resources. The obvious name for that stance is puritan. The obvious response to it is probably that people are entitled to follow that principle in their own lives, but not to impose it on others.

There are also values that academics are particularly likely to regard as having ethical force. The outstanding example is open access to information. That is reflected not only in attitudes to publication, but also in the issue of open source software. Both raise real conflicts with other parts of society.

It is entirely natural to wish that the research area were not beset with so many kinds of moral and ethical judgment, sophisticated or naive. But since the issues are there, it is better to see them clearly than to stumble through them in the dark.

3. Formal codes for affective computing

3.1 Ethics and human participants

The area where ethical guidelines are most formalised is human participation in research. That affects various roles that human participants play, such as providing material for databases, labelling it, and evaluating systems. Most of the activities are obviously harmless, but the fact that they are harmless still needs to be verified, and failure to do that can be disastrous. To complicate matters, the requirements vary with country, institution and application. A chapter like this cannot cover all of the variants, and the only

[1] http://www.epsrc.ac.uk/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx

safe rule is to check local regulations thoroughly before beginning research with human participants.

Concern about experiments with humans features in binding international agreements. For instance, the Charter of Fundamental Rights of the European Union[2] stipulates in Article 3 that the principle of free and informed consent must be respected. Strictly, it refers to medical and biological research, but 'biological' is routinely understood in a broad sense. For research that is classified as medical, there is a worldwide convention, the Declaration of Helsinki[3]. For non-medical research, the most highly developed codes are those that deal with psychological research. They have been used to govern experiments in affective computing (Sneddon et al 2011), and it is useful to give more detail about one.

The American Psychological Association (APA) code of ethics[4] begins by setting out principles, which are based broadly on Ross and Beauchamp and Childress. Those are followed up with detailed prescriptions for various aspects of the research process:

- Institutional Approval
- Informed Consent to Research
- Informed Consent for Recording Voices and Images in Research
- Client/Patient, Student, and Subordinate Research Participants
- Dispensing with Informed Consent for Research
- Offering Inducements for Research Participation
- Deception in Research
- Debriefing

The discussion of some of these issues is quite extensive, and it provides a useful starting point for anyone proposing to work with human participants.

Some systems are more restrictive than the APA code. For example, some ask for proof that the research will add to knowledge (if not, it should not be done). Medical protocols tend to be particularly exacting, because they are designed to deal with areas where both risks and rewards are very high. It is always worth checking whether that level of scrutiny is needed.

In general, approval is obtained by completing a form that sets out issues to be considered, and requires appropriate declarations on each. It is submitted to a research ethics committee. Specifications for the membership of the committee vary from place to place, but it is generally required to include 'lay members', meaning (roughly) that they have no professional connection with the research area. Specialised knowledge about ethics is rarely expected.

From an institution's point of view, the research committee has a dual function: to prevent harm, and to provide indemnity if harm is caused. Indemnity depends on agreement with the institution's insurers. Any group can constitute committees that carry out the first function, but they should be clear that the second is a different matter: if the experimenters are sued, the committee may simply end up sharing the bill.

3.2 Technological codes

Information technology has general professional codes analogous to the APA code above. A good example is provided by the Association for Computing Machinery (ACM) [5]. Like the APA code, it begins with relatively ethical standard principles, but it emphasises specific issues which are of particular concern to information technology. The most salient of those is privacy of data. It states:

> Computing and communication technology enables the collection and exchange of personal information on a scale unprecedented in the history of civilization. Thus there is increased potential for violating the privacy of individuals and groups. It is the responsibility of professionals to maintain the privacy and integrity of data describing individuals.

[2] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2010:083:0389:0403:en:PDF

[3] http://www.wma.net/en/30publications/10policies/b3/

[4] http://www.apa.org/topics/ethics/index.aspx

[5] www.acm.org./constitution/code.html

Like the rights of participants, the status of personal data is an internationally recognised issue. Article 8 of the EU Charter of Fundamental Rights states that “Everyone has the right to the protection of personal data concerning him or her”, and there is an ongoing process of articulating the implications of the principle. The results include legislation with strong implications for both the creation and the use of databases[6]. Data that individuates a person – which includes a great deal of material in affective databases – can only be collected “for specified, explicit and legitimate purposes”, and must not be “further processed in a way incompatible with those purposes”. It would seem that in a scientific context, the data may not be processed at all unless the subject of the data has unambiguously given his consent. There are very severe restrictions on the use of data revealing racial or ethnic origin, which both speech and photographs are likely to do.

These provisions are not static. For example, a recent EU report on Ethics of Information and Communication Technologies[7] described strengthened legislation on data protection, including a reinforced ‘right to be forgotten’ (people will be able to delete their data if there are no legitimate reasons for retaining it); and a requirement that consent for data to be processed will always have to be given explicitly, rather than assumed. It also explicitly recognised a range of other issues, including protection against cybercrime, restricting access for minors, and ensuring equality of access for groups who might be excluded. Some of these represent problems to be borne in mind, others quite possibly opportunities: for instance, removing affective barriers to computing may make a significant contribution to access (Cowie, 2012).

Although most of the report deals with well-trodden issues like confidentiality of data and equality of access, it picks out, largely in passing, a few concerns with more specific implications for affective computing – that the line between encountering reality and artificial surrogates may become blurred, and that vulnerable people may form undesirable attachments to artificial carers. The result is to bring those issues into a grey area between speculative discussion and binding codes. It means that at the very least, they are unsafe to ignore.

A different technological perspective is reflected in a position paper which was mentioned earlier, produced by the UK Research Council for Engineering and Physical Science (EPSRC)[8]. It is meant to govern research on robotics, but if one accepts the principles, then it would be hard to doubt that they should apply to virtual agents as well as robots. Five main principles are proposed:

1. Robots should not be designed as weapons, except for national security reasons.
2. Robots should be designed and operated to comply with existing law, including privacy.
3. Robots are products: as with other products, they should be designed to be safe and secure.
4. Robots are manufactured artefacts: the illusion of emotions and intent should not be used to exploit vulnerable users.
5. It should be possible to find out who is responsible for any robot.

These principles have obvious links to Asimov’s (1950) Laws of Robotics. Like Asimov, the authors are concerned with the risk of robots doing harm to humans, and they propose prohibiting it, even as a way of protecting valuable property (the robots themselves) from theft or vandalism – that is what lies behind point (3). Unlike Asimov, they oppose any blurring of the line between humans, who are responsible agents, and artefacts, however sophisticated. That makes ‘the illusion of emotions’ a subject of concern. Because of their concern about potential dangers, they also question what should be available as open code.

The EPSRC paper is in an interesting category. It is not binding, but it gives an official status to concerns that are widely held, and means that arguments to the contrary have to be carefully thought through. It also signals a point whose importance is hard to overstate. Formal codification is an ongoing process, and ignoring issues which have not yet been codified is a risky strategy.

[6] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:NOT

[7] http://ec.europa.eu/bepa/european-group-ethics/docs/publications/ict_final_22_february-adopted.pdf

[8] http://www.epsrc.ac.uk/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx

4. Ethical themes for affective computing

After the areas where there are quasi-legal codes, there are several well-established ethical themes with direct connections to affective computing. This section picks out five. They are not wholly independent. That is not carelessness – the fact that ethical principles do overlap is well known, and helps to make practical decision-making simpler than one might fear.

4.1 Benificence

This section is about a point which is both simple and fundamental to any balanced discussion. Affective computing is a technology with unusually direct links to morally positive goals. Its most obvious function is to make technology better able to furnish people with positive experiences, and/or less likely to impose negative ones. That means it has a direct relationship to what one major ethical tradition, utilitarianism, regards as the fundamental moral imperative: maximising net happiness. That general point is reflected in many specific efforts with morally positive goals: for examples, see Cowie (2012).

Once that is recognised, it is natural to separate two different types of objection to affective computing. On one side are concerns about unintended damage that might outweigh intended gains in net happiness – for instance, concerns about unintended effects on people involved in the research, or what the systems might do in the wrong hands. It is possible to engage with those. A different kind of difficulty arises when objectors deny that a shift towards positive affect has any moral value at all. In effect, they rule the natural positive out of court; and what is left is very likely to appear negative. For example, it is hardly surprising if people who start from a Kantian emphasis on autonomy and rationality see many risks and few gains. Less obviously, the same is true of those who argue that the happiness we should maximise is not positive affect (hedonia), but a subtler sense that our life is worthwhile (eudaimonia) (e.g. Ryan, Huta & Deci 2008).

'Positive psychology', which also values positive affect, encounters similar problems (Csikszentmihalyi & Csikszentmihalyi, 2006). Both have to reckon with people who see no moral value in shifting the balance between positive and negative affect. There are more serious things that should be occupying us. Logic cannot compel people to change that stance. However, if it is the basis on which they judge, they should be pressed to be clear about it, because by no means everyone agrees.

Two final implications should be drawn out. First, one of the obvious roles of affective computing is remedial. It is to spare people distress that would otherwise be caused by interactions with affectively incompetent systems (Cowie 2012, Scheutz 2012). Second, if we believe that affective computing could increase the net happiness of humanity, our ethical duty includes countering misguided fears that might prevent that; and, of course, ensuring that we do not inflame the fears.

4.2 Deception

Deception is widely recognised as a key problem area (Bringsjord & Clark 2012,Coeckelbergh 2012, Cowie 2012). It is not explicitly noted as an issue in most of the codes routinely invoked in ethics, but the implication that it is can be derived from the main codes. In terms of Ross's principles, it violates the duty of fidelity (to keep our promises). In terms of Principalism, it infringes autonomy, because misinforming a person about the alternatives that are open prevents him or her from choosing rationally between them. The Golden Rule will appear shortly.

There are various more technical discussions of computing and deception. It features in the ACM guidelines. The literature on persuasion includes well-known guidelines (Berdichevsky & Neuenschwander 1999). There is also a more recent, high profile literature on specifically emotional misdirection, usually in the context of artificial companions and carers. The sources take quite different approaches, which reflect deeper divisions in the ways that deception might come about.

It is useful to begin with the most general charge. It is said that the whole enterprise of affective computing is deceptive, and cannot avoid being deceptive – and therefore, the whole enterprise is unethical. According to that argument, core parts of it rest on giving the impression that systems feel emotions when in reality, they have no feelings of any sort (Sparrow 2002, Coeckelbergh 2012).

That kind of claim strays into questionable territory. It assumes that people normally treat emotion-related signals as declarations that some internal feeling state exists. That seems unlikely. Certainly people do not usually regard it as dishonest to give signs of a positive feeling that does not exist ('whistling in the dark'). Nor do they deplore artefacts which display signs of emotions that they do not have – paintings, movie images, dolls, and so on. Animals show signs that people are highly disposed to read as signalling emotion, but very few people are deeply concerned by the question of what feeling state, if any, goes with them. The point here is that whatever the concerns here are, they need to be framed in a way that does not suggest we should be equally concerned about optimists, animals, dolls, and the Mona Lisa.

At the other extreme, it seems plain that systems should not be deliberately engineered to make people believe something that is actually false. Emotional competence certainly can enhance the ability to deceive, and for that reason it is natural to fear that if we start with the standard mechanical virtues of flawless logic, endless patience, and no conscience, and add the ability to manipulate emotion, the result could be an almost irresistible persuader (Guerini & Stock 2005). There would clearly be ethical objections if a system of that kind were used, for instance, to convince people that they should buy a financial service which was actually inappropriate.

A well-known discussion of 'persuasive technologies' by Berdichevsky & Neuenschwander (1999) addresses that issue. It proposes a guideline based on the Golden Rule: "The creators of a persuasive technology should never seek to persuade anyone of something they themselves would not consent to be persuaded of." Various refinements of the approach have been proposed (e.g. Spahn 2012), but it is clear that a principle of that general kind is ethically important.

The two concerns considered so far involve extremes – inevitable deception and deliberate deception. Between them is a difficult grey area. Two principles may help to separate legitimate worries from overstatement. First, people may not worry greatly about signs that do not truly reflect internal feeling states, but they do object if the signs mislead them about the way a system is likely to behave – particularly if the false impression affects their own choice of action. Second, signs of emotion are prone to create a particular kind of false impression, even when no outright deception occurs. The problem involves what has been called 'pars pro toto' reasoning (Cowie, 2012). It occurs when a system shows some behaviours associated with an emotion, and people infer that it has a complex of other characteristics that would be associated with that emotion in a human. In a sense, people who form that kind of impression are deceiving themselves; but it is such a characteristically human type of inference that the system designers can hardly disclaim responsibility.

The obvious illustration is where an agent acting as a teacher or a companion uses facial and vocal gestures that give an impression of caring. That may help the agent in its intended function, but it is a problem if the user drifts into assuming that it will show other kinds of caring behaviour, and relies on it for help that it cannot actually provide. Teacher and companion roles are mentioned because it is a problem that we might expect to be worst where users did not have full adult judgment.

There is no straightforward way to forestall that kind of problem. The obvious prescription is that users and/or their representatives should be involved in identifying possible misinterpretations at the design stage.

An important complication in this area is that it is not clear how far ethical responsibility goes back. In particular, what ethical responsibility attaches to a research team who designed a basic system with no intention to deceive, but who did nothing to prevent it from being customised to deceive? The issue here is closely related to the 'open source' clause in the EPSRC code that was mentioned in earlier sections. The code highlights the likely outcome: both the law and the public would probably hold the basic research team responsible unless they had taken active steps to prevent foreseeable abuse.

4.3 Respect for autonomy

A third widely recognised ethical theme which applies to affective computing is autonomy. Respect for autonomy is widely regarded as fundamental to a liberal society (e.g. Dworkin,1988). That is partly because the implications that can be derived from the principle go much further than one might immediately realise. At least some of the implications clearly raise questions for affective technology.

Central to the implications is the notion of procedural independence. People have the potential for autonomy, but to exercise it, they must have procedural independence – that is, freedom from factors that compromise or subvert their ability to achieve self-refection and decide rationally (Dworkin 1988). The potential to infringe procedural independence is an issue in various ways.

One has already been mentioned in the discussion of deception. Deception violates two kinds of ethical principle, a duty of honesty in and of itself, and a duty not to infringe autonomy. The reasoning is that giving people a misleading impression of the alternatives impairs their procedural independence, and thereby their ability to make rational decisions. That kind of impairment is not always a pressing issue, but when it is, deception is doubly unethical.

A second line of argument on autonomy has been developed by Baumann & Doring (2011). It turns on agents' ability to perceive emotion rather than to persuade. They argue that information about a person's emotional state has particular implications for procedural independence: if it becomes available, it can restrict their options in ways that they would not choose. Hence they propose two duties with respect to information about other people's emotional states:

> First, persons should respect other persons' control of access to information about their emotional states.
>
> Second, the fact that persons obtain or are entrusted with knowledge about emotional states of a person imposes special responsibilities upon them: They must not misuse the information and exploit the vulnerabilities of that person.

As a result, people need strong assurances about the use that will be made of any information that an artificial system obtains about their emotional states. The governing principle that they propose is that

> Emotion-oriented systems should not undertake any actions that users – as autonomous persons – do not or cannot endorse.

where the primary kind of action being considered is use of information about the person's emotional state. Clearly, this is a variant on concerns about access to information that were raised earlier. But as with deception, the concern may have a double force when the information in question is about emotions.

Issues in a third area are linked via the notion of respect. Respect for autonomy is at least often understood to mean that beings capable of autonomy have a unique status, which is owed respect; and their autonomy should not be threatened by undermining their self-respect. For that reason, communication which denies respect is ethically problematic. The implications are wide-reaching. One which is at least beginning to be explored is that agents should respect conventions of politeness (Brunet et al 2012). On the standard account (Brown & Levinson 1987), the function of politeness is to avoid threatening the other person's 'face'. Hence impolite communication does not simply violate conventions: it denies respect, and threatens self-respect.

Affective computing is the technology best placed to develop polite communication. Most of the signals that are used to express emotion – smiles, nods, and postures as well as selected forms of verbal expression – have a key role in politeness; and the point is to affect people's feelings in particular ways (or to avoid affecting them). The ethics are not straightforward. Some forms of politeness may lead to misunderstandings (Bonnefon et al, 2011), and so there can be tension between truthfulness and according respect. Nevertheless, it seems right to insist that there are good ethical reasons to explore ways of incorporating some functions of politeness into human-computer interactions.

4.4 Certifying competence

The ACM code cited above includes a responsibility to “ give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.” One of the ethical problems that arises with affective computing is that it is extremely difficult to discharge that responsibility. It is well known that evaluation of affective systems is problematic (e.g. Schroeder et al 2011). The problem is not avoidable, because the computer systems’ function is intrinsically bound up with human systems that we understand only very partially. It is usually impossible to guarantee analyses of risk, partly because the human systems are very complicated, and partly because they are very incompletely understood. Added to those is an understandable reluctance to proclaim the limitations of a product that represents an enormous investment of effort and intelligence. Nevertheless, failure to proclaim them is a real ethical problem.

The most sustained discussions of the issue in the context of SI IFs, that is, Semi-Intelligent Information Filters (Goldie et al 2011, Cowie 2012). These are supposed to detect practically important emotion-related states, and to pass their conclusions on for action of some sort. The danger is that they will have limitations that are poorly understood by those who deploy them, and as a result, people will be subjected to actions that they do not deserve, or will not receive responses that they ought to. The problem is not new. The classical example involves ‘lie detectors’. Despite widespread belief in their powers, they were actually much more likely to stigmatise the innocent than to pinpoint the guilty (National Research Council, 2003). That experience could easily be repeated in areas such as surveillance, monitoring employees, detecting distress in phone calls, and so on. There are overwhelming reasons, both practical and ethical, to avoid that.

The problem is approached from a different angle by Sloman (2010). He stresses the obligation to analyse fully what a function entails before we claim that a system can carry it out. His context is a discussion of artificial helpers, and he provides a daunting overview of the abilities that a system would need to be a competent helper. However, the principle generalises. Before we think of claiming that a system is empathetic (Janssen 2012), let alone loving, we need analyses of what those functions entail against which we can measure what the system does.

4.5 Portraying humans

As noted earlier, the way ethical principles are usually formulated makes it natural to focus on what systems do. However, a different type of ethical issue features too often to ignore. Broadly speaking, it involves the way affective computing portrays human beings. Its portrayals raise ethical issues because of the intimate connection between emotion and morality that was also noted early in the chapter. The issue arises at two main levels.

The more concrete level stems from the fact that in everyday language, descriptions of emotion are rarely morally neutral. Hence, to use them is to pass a kind of moral judgment; and it is not obvious when a machine has the right to pass that kind of judgment. Cowie (2005) noted an extreme case. To say that a person is sulking is to pass a moral judgment, and it is hard to imagine people accepting that a machine had the right to do that. A subtler case involves a machine recording that a person is angry, but not what made them angry. That means the output provides no way to make the key moral judgment about anger, which is whether it is justified. If the default assumption is (as seems likely) that anger is not justified, then a person who felt that their anger was justified might have grounds to take exception.

A subtler issue involves presenting phenomena that are morally entitled to certain kinds of human response – typically empathic – in a way that disguises that kind of significance. For example, there is considerable interest in systems that detect pain and distress (e.g. Lucey et al 2009, Roberts 2010). Natural methods of detection involve responding to facial expressions and attributes of speech that evoke empathic as well as diagnostic responses. The diagnosis may be better if the situation is portrayed by a line on a graph showing levels of pain or distress, but eliminating the empathic elements is not a trivial matter.

The issues here are quite intricate. It might be argued, by analogy with the laws of libel, that there is no cause for concern unless the descriptions are used in ways that harm the person described. However, it is also intuitive to say that people have an obligation to recognise that they are simply not entitled to pass certain kinds of judgment, even if they keep it private. We might expect the same to hold for a machine, if indeed it is entitled to pass any kind of moral judgment. The main point is simple, though. People who deal in morally sensitive concepts have an obligation to recognise the moral issues that the concepts raise.

Moving to the more abstract level, one of the oldest debates in philosophy is how we should value parts of our makeup other than pure intellect. For the stoics, the right way to live (and therefore the morally proper goal) was to acheive *apatheia,* where emotion was completely subordinated to intellect. Augustine retorted that those who are "not stirred or excited by any emotions at all ... lose every shred of humanity" (p.566). The sketch of ethical traditions at the beginning of this chapter shows that the tension continues.

Affective computing cannot separate itself from that debate, because it affects our understanding of the systems underlying human emotion – particularly, how crude or sophisticated they are. A good deal of research argues that when we try to match what emotion-related processes achieve, what we find is that they are vastly more impressive than the unaided intellect usually recognises (Cowie, 2009). That favours Augustine. However, on the other side, there is literature that suggests emotions are nothing more than heuristics that can be incorporated in a toy dog (Aibo is said to have 'real emotions and instincts'[9]), or a set of numbers that rise under certain eliciting conditions and decay in a certain temporal pattern (Bryson & Tanguy 2010). The difference is ethically important because the weight we should attach to emotional reactions is, and has been for millennia, a central question in ethics.

A related, but distinct issue arises for those who believe in the intrinsic value of life (see, e.g. Link, 2013). From that viewpoint, it is morally disturbing to propose that there is any meaningful correspondence between a manifestly lifeless system, and something as central to life as emotion. It is asserting that one of the things people value most deeply is, in reality, much like something that they do not value at all.

As before, the issues are intricate, but they point to a simple obligation. The way affective computing portrays emotion has far-reaching implications for the way people understand themselves. It may in the long run help to resolve ancient questions. In the meantime, people who work in the area have a moral obligation to recognise how sensitive their pronouncements are.

4.6 Application-specific concerns

The themes that have been sketched so far were chosen partly because they subsume most of the ethical issues in a wide range of application areas. Applications that seem to be reasonably well covered include games, advertising, presenting a corporate image, instruction, and non-medical coaching or training. However, that leaves several areas that raise more specific issues.

Affective systems as companions There is heated controversy over the use of affective systems as 'companions'– most notably as part of a caring role for the elderly, but also for children (Sharkey 2008, Wilks 2010). Previous sections have already covered concerns that are directly related to the contribution of affective computing, notably in the context of deception and certifying competence. However, if affective computing is involved, it also has to register ethical concerns that affect the whole enterprise. At a general level are concerns that "the seductions of the robotic" may lead people "to sidestep encounters with friends and family" (Turkle 2010 p.7), or that 'the robotic' allows friends and family to sidestep their responsibilities. Others are more technical, such Newell's (2010) observation that providers owe users a duty to find out what they actually want. There is an ethical obligation on people who venture into the area to be informed about issues like these.

Affective systems in medicine Affective computing has a growing range of applications related to medicine. Functions include counselling (Marsella et al 2000), psychological therapy (Kang et al, 2012), and aids to diagnosis (Ashraf et al 2009, Trevino et al 2011). The outstanding issue in medical applications is the potential for extreme consequences when a therapy goes wrong, well reflected in the title of the standard text on medical ethics, "Causing Death and Saving Lives" (Glover, 1977). That is reflected in the particular

[9] http://www.robotbooks.com/sony_aibo.htm, downloaded 28.2.2013

thoroughness of ethical approval procedures in medical contexts, which has already been pointed out. Glover's book is an excellent introduction for anyone considering work in the area.

Military Affective computing does not immediately suggest military applications, but in fact, some of the longest-established work in the area is concerned with detecting stress in combat situations (e.g. Vloeberghs et al 2000). The EPSRC report cited earlier rejects the development of robots designed "to be used as weapons with deadly or other offensive capability" except for national security, but many would argue against any development of killing machines – literally – that might escape trustworthy control (Sharkey, 2008). Other issues involve concern that terrorists might acquire potentially deadly technologies (again, see the EPSRC report), and hence the ethic of open publication . Again, the main point is that people considering research with military connections should properly weigh the issues.

Sex robots Satisfying sexual fantasies is one of the most potentially lucrative applications of affective computing. Widespread moral codes regard that as a thoroughly unethical activity. However, a recent paper on the subject (Sullins 2012) concludes that "the attainment of erotic wisdom is an ethically sound goal" (p.398) provided that the system respects limits on the manipulation of human psychology. There are few clearer examples of the point that there are ethical differences which pure logic will not resolve.

Surveillance Some ethical positions that imply surveillance is almost always wrong, because it infringes autonomy. For others, it is likely to promote happiness more often than distress, and therefore profoundly moral. As with military applications and sex robots, that is an issue that logic will not settle.

It should be noted that at this level, different cultures differ quite markedly – for example, there are very different tolerances for surveillance or use of robots in traditionally human roles. Although that is widely accepted, systematic research on the topic is in early stages. [10]

5. The enforcement of ethical principles and concerns

The discussion so far has focused on what should and should not happen. This section considers how compliance is enforced, and how that impacts the research process. Because different countries and cultures have different systems, what is said here can only be a starting point for someone working in a particular setting. However, nobody should forget that some constraints cut across local systems. An institution may allow research to go ahead without ethical scrutiny, only for the research team to discover that a journal will not publish their work because there is no ethical documentation.

It makes sense to begin with the effort to fund research. Funding bodies have very diverse procedures, but the process of deciding whether to fund a project tends to involve several levels of test. The European Union's Framework 7 program (FP7) can be taken as an example.

The program's rules underline the importance of ethics. They state that "any proposal which contravenes fundamental ethical principles ... shall not be selected and may be excluded from the evaluation, selection and award procedures at any time." [11] As a first stage, applicants are routinely asked to complete a checklist of descriptions that are associated with well known problems. The FP7 list includes the following items that are potentially relevant to affective computing:

Informed Consent

- Does the proposal involve children?
- Does the proposal involve patients or persons not able to give consent?
- Does the proposal involve adult healthy volunteers?
- Does the proposal involve Human Genetic Material?

[10] http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/G069808/1
[11] ftp://ftp.cordis.europa.eu/pub/fp7/docs/guidelines-annex5ict.pdf

- Does the proposal involve Human biological samples?
- Does the proposal involve Human data collection?

Privacy

- Does the proposal involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)
- Does the proposal involve tracking the location or observation of people?

Dual Use

- Research having direct military application
- Research having the potential for terrorist abuse

It is noticeable that the list covers only issues that are well-established, and apply to a wide range of research areas. It does not touch on most of the issues that have been discussed up to this point. However, it is only the first step. The checklist needs to be followed by text explaining how the research will handle issues arising from the list, and any others. Checklist and text are considered in the evaluation process, and the last of five sections in the evaluators' report asks whether the proposal raises "ethical issues that need further attention?" If it does, a specialist ethical review may be called for.

Specialist ethical reviews, and to a lesser extent panels, are informed by expert groups – in the case of FP7, the European Group on Ethics in Science and New Technologies[12]. It produced the report on Ethics of Information and Communication Technologies which was cited earlier. That system means that issues can become important in review panels' deliberations very quickly.

The EU is by no means isolated in that approach. For example, at the time of writing, the UK's Engineering and Physical Science Research Council is sponsoring the development of a framework for ethics in ICT, based on "a comprehensive baseline study of current issues, challenges and responses to them as perceived by ICT researchers". [13] It is the norm for funding bodies to regard ethical use of technology as a moving target, and reasonably so.

Institutional scrutiny was considered early on in the chapter. The main point to be made here is a contrast. Whereas funding bodies will typically consider the broad outline of a proposal, local bodies may be very exercised by details like the exact wording of a consent form. Addressing details may take several iterations, and that may be a real problem if the committee only meets half a dozen times a year.

Once research has been done, another set of ethical filters applies at publication. The journal PLoS ONE is a useful example. [14] Its guidelines specify that if research has used human subjects, the method section must cover the following:

- the approving institutional review board or equivalent committee(s);
- how informed consent was obtained;
- if humans were categorized, how that was done;
- if potentially identifying material is published, explicit consent from the individual(s) concerned.

For observational or field studies, there must be ethics statements that specify the permits and approvals obtained for the work, including the authority that approved the study. In addition, "outmoded terms and potentially stigmatizing labels should be changed to more current, acceptable terminology". Papers that fail these tests are returned without review.

Beyond formal sanctions, it is striking how often media reports focus on perceived ethical problems rather than technical achievements. For example, it was a surprise when a report following a recent interview on the recognition of natural speech concluded:

> Unfortunately, if computers do ever get to the point where they can understand our words and how we say them, it might not be a good thing. There is a hypothesised crisis [the 'uncanny valley'] where interactions between humans and robots reach a level of realism that is uncomfortable and

[12] http://ec.europa.eu/bepa/european-group-ethics/index_en.htm

[13] http://gow.epsrc.ac.uk/Search.aspx?search=ethics

[14] http://www.plosone.org/static/guidelines.action#human

> disconcerting. ... So perhaps the question should be less about when we will be able to create computers that can draw on all the experiences and knowledge amassed over a lifetime when having a conversation, and more on whether we should be heading down this route at all.

Negative presentations in the media are not sanctions in themselves, but the risk that they will translate into sanctions is all too real. They shape public opinion, and public opinion sways funding bodies, particularly when the issue is perceived to be ethical: no politician wants to be held responsible for funding Dr. Frankenstein. Public opinion has made government funding for research on GM crops an extremely delicate issue (UK Parliamentary Office of Science & Technology 2012), and virtually ended some kinds of research with animals. Avoiding the same fate is not a trivial matter.

6. Public intuitions and duties to explain

The discussion so far has emphasised links between ongoing research and well-established ethical principles. In general, though, the issues that preoccupy the public are different. They are usually more to do with gut feeling than with arguments based either on ethical principles or on knowledge of affective computing. However, as the first section noted, the gut matters in ethics. It is ethically suspect as well as risky to persist with activities that genuinely perturb the public. The general response that this section proposes is not to abandon the activities, though. It is to accept an ethical duty to ensure that the public can form rational judgments.

6.1 The ethics of unquantifiable risk

An area where there is some clarity involves risk. The problem, as pointed out by Goldie et al. (2011), is that the risks involved are profoundly unquantifiable. Certainly the outcomes that people fear are deeply disturbing. On the other hand, there is no convincing case for thinking that they are likely. But yet again, they cannot be ruled out either.

One issue of that kind has already been mentioned. It is that surrogate worlds may become so engaging that people lose the will, and perhaps the ability, to relate to the real one. The EU, for example, regards that as a substantial issue. The ethically sound reply would seem to be that it does make sense to monitor the issue, but also to make it clear that there is no obvious reason to expect effects which are either particularly pernicious or particularly difficult to control.

Even more disturbing is the fear that 'mainds' will outstrip humanity and – at best – subordinate it. Affective computing has a special place in that nightmare because it raises the prospect of machines that can decide, not necessarily rationally, what they like and dislike. If that were a realistic possibility, then people ought to worry about it. The ethical issue here would seem to be helping non-experts to gauge the probability. If people's fears are unnecessary, then those who know the reality have an ethical obligation to avoid inflaming it, and if possible, to reduce it by exposing the limits of the machines that can actually be built or envisaged.

6.2 The ethical status of an agent

Concepts like 'autonomy' and 'free will' loom large in ethics (particularly Kant's), and also in the public mind. In both cases, what they mean is not simply that an agent can operate without being told what to do at every choice point. It is that the choice is fundamentally its own, rather than a product of various (probably ill-defined) background influences. For Kant in particular, an agent cannot make an ethical choice unless it has that kind of freedom to begin with.

One of the recurring concerns about affective computing is that to give systems true emotions would be to give them that kind of autonomy: and to give them true autonomy is to court disaster. The EPSRC report which has been cited repeatedly illustrates various concerns of that kind. Part of it is that autonomy should not be taken out of human hands; part is that machines are not actually capable of it; part is that giving them autonomy would distance the machines' human makers from responsibility for their actions; part of it is that machines with that kind of autonomy might turn on their human makers.

Concerns like that brings into play one of the oldest ideas in philosophy (expressed in Plato's image of the charioteer): that systems capable of initiating action, of which emotional systems are a prime example, need to be under rational control. The interplay between rational control and emotion-driven 'action tendencies' (Frijda, 1987) is a central theme in ethics, and there are contrasting views of the way it could or should play out in artificial systems.

Perhaps the simplest view has been put forward by Beavers (2009). For him, the reason/emotion tension is a human phenomenon, which should not be imported into artificial agents. Their decision-making can and should be wholly rational – which, on his Kantian view, would mean that it would not be moral at all. If so, morality has no place in artificial systems. In response, Guarini (2012) has argued that practically, it is likely that conflicts like humans' will arise. If so, artificial emotion and artificial ethics need to go hand in hand. Influential models suggest an even tighter connection. It has long been recognised that moral judgments are part and parcel of at least some emotions – for example, righteous anger (Plato's example) and remorse. Recently there has been growing interest in the connections between morality and emotions that seem at first sight purely biological, notably disgust (Schnall et al 2008, Erskine et al 2011). If so, attempts to model emotion apart from morality are misguided. More radical still is the proposal that empathy, which is primarily affective, not rational, lies at the root of moral behaviour towards others (Baron-Cohen, 2012). It does not follow automatically that systems which lack human-like emotions cannot behave ethically towards human beings, but it is certainly an issue to ponder.

Intellectually, these issues are fascinating. However, it bears emphasis that they belong in the realms of speculation, not practice, because the systems that we can currently build have, by human standards, very few courses of action to choose between. So long as we build systems to do only a few things, and specify when they should do which, it is hard to dispute the EPSRC conclusion that attributing ethical responsibility to them simply clouds the issue: responsibility lies firmly with the builder.

6.3 Mysterious forebodings

It cannot be proved, but it is a fair guess that a great many ethical arguments draw sustenance from reactions with at least echoes of the supernatural. There are two obvious kinds of reaction in that category.

One is revulsion at things that approach naturalness, but miss it in critical ways. The effect is usually described using Mori's (1970) phrase the 'uncanny valley', suggesting a fall in acceptability that sets in when things that are not natural creatures become too similar to them. There is a strong tendency to attach ethical significance to the uncanny valley, as the press report cited above illustrates: it interprets the effect as a reason to question whether the research should be done.

The effect is not as inevitable as Mori's description suggests (MacDorman, 2006). Nevertheless, the horror industry testifies to the strength of human reaction to some things that are human-like, but not human. Clearly that kind of effect raises practical issues – systems that people find profoundly disturbing will fail, because people will not use them. It would also raise ethical issues if people were forced to interact with systems which had that effect on them, and that might happen if, for instance, the systems were operating in a hospital or care environment.

It is another matter, though, to move from disconcerting experiences with a system to the conclusion that it, and the enterprise that produced it, were somehow evil. Since ethics presupposes rationality, from an ethical point of view, the appropriate response would presumably be that that was superstition, and should be resisted.

Similar in some senses, contrasting in others, is the sense that there is ground where humans should not tread. Some things are too mysterious to tamper with, and if we do tamper, the consequences may be wildly unpredictable. The concept of artificial minds is certainly surrounded by that kind of thinking. For instance, in a recent science fiction novel, true artificial intelligence is brought into being accidentally when a schoolgirl uploads data on her pet rat's brain (Brin, 2012). What unfolds then depends more on the rat than on the humans. Ideas about emotion have a similar quality – it is natural to feel that once we coax the glowing fluid into the circuitry, we have released forces beyond our control.

If it were reasonable to believe that emotion was like that, then it would be ethical to oppose affective computing on the grounds that it was profoundly dangerous. However, the belief is not reasonable, and the ethical course is surely to explain to people who are troubled by that kind of image why it is not reasonable.

7. Conclusion

This chapter has covered a wide range of issues. The nature of the area makes that inevitable. Perhaps surprisingly, though, it converges on a reasonably short list of ethical obligations that anyone who worked in affective computing should respect.

- They should understand the premises on which ethical judgements are likely to be based, and the fact that others may rationally hold ethical premises different from their own.
- They should abide by the ethically motivated codes that govern studies with human beings and data privacy.
- They should uphold the ethical value of making interactions involving humans and machines more likely to generate positive affect and less likely to generate negative affect.
- They should seek to ensure that the systems they build will do nothing to others that they would not want to be subjected to themselves, and nothing that users would object to if they understood what was happening.
  - In particular, they should ensure that their systems do not deceive people or infringe their autonomy in ways that violate that principle
- They should ensure that they have a clear understanding of the capabilities and limitations of their systems, grounded in an understanding of the corresponding human capabilities.
- Their communications with non-experts should help them to form realistic assessments, both of the systems' abilities, and of the risks that they might pose.
- They should be sensitive to the moral implications attached to terms that they use and models that they propose.
- Where the fields in which they work raise other ethical issues, they should become familiar with them.

It would be completely against the spirit of the chapter to expect instant agreement on that kind of list. However, it seems reasonable to offer it as a point of reference.

8. References

Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., & Solomon, P. E. (2009). The painful face–Pain expression recognition using active appearance models. *Image and Vision Computing*, 27(12), 1788-1796.

Asimov, I (1950) *I Robot* New York: Doubleday

Augustine (Tr Bettenson) (1984) *City of God* London: Penguin

Baron-Cohen, S (2012) *Zero Degrees of Empathy: A New Theory of Human Cruelty and Kindness* London: Penguin

Baumann, H. and Döring, S. Emotion-Oriented Systems and the Autonomy of Persons In P. Petta · C. Pelachaud · R. Cowie (Eds) *Emotion-Oriented Systems: The Humaine Handbook* Berlin: Springer pp. 735-752

Beauchamp, T.L. and Childress, J.F. (2001) *Principles of Biomedical Ethics, 5th edn*. New York: Oxford University Press.

Beavers, A. (2009) Between Angels and Animals: The Question of Robot Ethics, or Is Kantian Moral Agency Desirable? *Proc. Assoc. For Practical and Professional Ethics 18th Ann. Meeting* http://faculty.evansville.edu/tb2/PDFs/Robot%20Ethics%20-%20APPE.pdf downloaded 28.2.2013

Berdichevsky, D., & Neuenschwander, E. (1999). Toward an ethics of persuasive technology. *Communications of the ACM*, 42(5), 51-58.

Blackburn, S. (2001) *Ethics: A Very Short Introduction*. Oxford: Oxford University Press.

Bonnefon, J. F., Feeney, A., & De Neys, W. (2011). The risk of polite misunderstandings. *Current Directions in Psychological Science*, 20(5), 321-324.

Brin, D 2012 *Existence* London: Orbit Books

Bringsjord, S. & Clark, M. H. (2012) Red-Pill Robots Only, Please *IEEE Transactions on Affective Computing* 3,394-397

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* Cambridge: Cambridge University Press.

Brunet, P. M., Cowie, R., Donnan, H., & Douglas-Cowie, E. (2012). Politeness and social signals. *Cognitive Processing* 13 S447-453.

Bryson, J. J., & Tanguy, E. (2010). Simplifying the design of human-like behaviour: Emotions as durative dynamic state for action selection. *International Journal of Synthetic Emotions*, 1(1), 30-50.

Coeckelbergh, M. (2102) Are Emotional Robots Deceptive? *IEEE Transactions on Affective Computing* 3, 388-393.

Cowie, R. (2005). What are people doing when they assign everyday emotion terms? *Psychological Inquiry*, 16(1), 11-48.

Cowie, R. (2009). Perceiving emotion: towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), 3515-3525.

Cowie, R. (2012). The good our field can hope to do, the harm it should avoid. *IEEE Transactions on Affective Computing* 3, 410-423

Csikszentmihalyi, M., & Csikszentmihalyi, I. S. (Eds.). (2006). *A life worth living: Contributions to positive psychology.* New York: Oxford University Press.

Döring, S., Goldie, P. and McGuinness, S. (2011) Principalism: A Method for the Ethics of Emotion-Oriented Machines . In P. Petta , C. Pelachaud & R. Cowie (Eds) *Emotion-Oriented Systems: The Humaine Handbook* Berlin: Springer pp. 713-724

Driver, J. (2012) *Consequentialism* Abingdon: Routledge

Dworkin G (1988) *The theory and practice of autonomy* Cambridge, MA: Cambridge University Press

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A Bad Taste in the Mouth Gustatory Disgust Influences Moral Judgment. *Psychological Science*, 22(3), 295-299.

Foot, Philippa 2001 *Natural Goodness* Oxford: Oxford University Press

Frijda, N. H. (1987). *The Emotions* Cambridge: Cambridge University Press.

Glover, J. (1977) *Causing Death and Saving Lives* London: Pelican Books

Goldie, P., Döring, S. and Cowie, R. (2011) The Ethical Distinctiveness of Emotion-Oriented Technology: Four Long-Term Issues In P. Petta · C. Pelachaud · R. Cowie (Eds) *Emotion-Oriented Systems: The Humaine Handbook* Berlin: Springer pp. 725-733.

Guerini, M., & Stock, O. (2005). Toward ethical persuasive agents. In *Proceedings of the International Joint Conference of Artificial Intelligence Workshop on Computational Models of Natural Argument*. July 2005.

Guarini, M. (2012) Conative Dimensions of Machine Ethics: A Defense of Duty *IEEE Transactions on Affective Computing* 3,434-442

Hobbes, T (1651/1996) *Leviathan* ed J.C.A. Gaskin Oxford: Oxford World Classics

Hume, D. (1740/2007) *A Treatise of Human Nature*: A Critical Edition (eds.) D. F. Norton and M. J. Norton, Oxford: Clarendon Press.

Janssen, J. H. (2012). A three-component framework for empathic technologies to augment human interaction. *Journal on Multimodal User Interfaces*, 5, 143-161.

Kang, S. H., Gratch, J., Sidner, C., Artstein, R., Huang, L., & Morency, L. P. (2012). Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* Volume 1 pp. 63-70

Lau, H. (2012) The rise and fall of voice Institute of Physics http://www.physics.org/featuredetail.asp?id=76

Link. H.J. (2013) Playing God and the Intrinsic Value of Life: Moral Problems for Synthetic Biology? Science and Engineering Ethics 19(2), pp 435-448

Lucey, P., Cohn, J., Lucey, S., Matthews, I., Sridharan, S., & Prkachin, K. M. (2009) Automatically detecting pain using facial actions. In *Affective Computing and Intelligent Interaction 2009*. (pp. 1-8)

MacDorman, K. F. (2006). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science* pp. 26-29

MacIntyre, Alasdair, 1985, *After Virtue*, London: Duckworth, 2nd Edition

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong* London: Penguin

Marsella, S., Johnson, W. L., & LaBore, C. (2003, July). Interactive pedagogical drama for health interventions. In *Conference on Artificial Intelligence in Education, Sydney, Australia*. http://alelo.co.uk/files/AIED03-interactive_pedagogical.pdf downloaded 28.2.2013

Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33-35.

National Research Council Committee to Review the Scientific Evidence on the Polygraph (2003) The *Polygraph and Lie Detection*. http://www.nap.edu/openbook.php?record_id=10420 downloaded 28.2.2013

Newell, A. (2010) Artificial Companions in society: Consulting the users. In Y. Wilks (ed) *Close Engagements with Artificial Companions* Philadelphia: John Benjamins pp 173-178

Parfitt, D (2011) *On What Matters*, Oxford: Oxford University Press

Roberts, L. (2010). Real and acted responses of distress: an auditory & acoustic analysis of extreme stress & emotion. *ExLing* 2010, 149-152.

Ross, W.D. (1939) *Foundations of Ethics*, Oxford: Oxford University Press,

Ryan, R. M., Huta, V., & Deci, E. L. (2008). Living well: A self-determination theory perspective on eudaimonia. *Journal of Happiness Studies*, 9(1), 139-170.

Scanlon, T. M. (1998) *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.

Scheutz 2012 The Affect Dilemma for Artificial Agents: Should We Develop Affective Artificial Agents? *IEEE Transactions on Affective Computing* 3,424-433

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34(8), 1096-1109.

Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D.& Wollmer, M. (2012). Building autonomous sensitive artificial listeners., *IEEE Transactions on Affective Computing* 3(2), 165-183.

Sharkey, N. (2008) The ethical frontiers of robotics *Science* Vol. 322 (5909), 1800 - 1801

Sloman, A. (2010) Requirements for artificial companions: it's harder than you think In Y. Wilks (ed) *Close Engagements with Artificial Companions* Philadelphia: John Benjamins pp 179-200.

Smith, A. (1759) *The Theory of Moral Sentiments* Printed for A. Millar, in the Strand; and A. Kincaid and J. Bell, in Edinburgh

Sneddon, I., Goldie, P. and Petta, P. (2011) Ethics in Emotion-Oriented Systems: The Challenges for an Ethics Committee In P. Petta · C. Pelachaud · R. Cowie (Eds) *Emotion-Oriented Systems: The Humaine Handbook* Berlin: Springer pp. 753-767

Spahn, A.(2011) And Lead Us (Not) into Persuasion…? Persuasive Technology and the Ethics of Communication. *Science and Engineering Ethics* (Published online first May 5, 2011), doi: 10.1007/s11948-011-9278-y

Sparrow, R. (2002). The march of the robot dogs. *Ethics and information Technology*, 4(4), 305-318.

Sullins, J. (2012). Robots, Love and Sex: The Ethics of Building a Love Machine. *IEEE Transactions on Affective Computing* 3, 398-409

Trevino, A. C., Quatieri, T. F., & Malyska, N. (2011). Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, 2011(1), 1-18.

Turkle, S. (2010) In good company? On the threshold of robotic companions In Y. Wilks (ed) *Close Engagements with Artificial Companions* Philadelphia: John Benjamins pp 3-10

UK Parliamentary Office of Science & Technology (2012) GM in Agricultural Development *Postnote Number 412* June 2012

Vloeberghs, C., Verlinde, P.,Swail, C., Steeneken, H., South, A. (2000) *The Impact of Speech Under "Stress" on Military Speech Technology* NATO Research and Technology Organization Technical Report ADA377422

Wood, A. (1999). *Kant's Ethical Thought*. Cambridge: Cambridge University Press.