humaine

D10b

Interim report to plenary meeting on ethical frameworks for emotion-oriented systems 2005

Peter Goldie, Sabine A. Döring and WP10 members

Version 1.0

Date: 29th June 2005

IST project contract no. **Project title**	507422 **HUMAINE** **Human-Machine Interaction Network on Emotions**
Contractual date of delivery **Actual date of delivery** **Deliverable number**	*month 17* *29 June 2005* *D10b*
Deliverable title	*Interim report to plenary meeting on ethical frameworks for emotion-oriented systems 2005*
Type	Public Report
Number of pages	14
WP contributing to the deliverable	WP10
Task responsible	KCL
Author(s)	Peter Goldie, Sabine A. Döring and WP10 members
EC Project Officer	Philippe Gelin

Address of lead authors:
Peter Goldie and Sabine A. Döring
Department of Philosophy
King's College London
Strand
London
WC2R 2LS
peter.goldie@kcl.ac.uk
sabine.doering@kcl.ac.uk

Authors also contributing:
Roddy Cowie, Ian Sneddon (QUB), Paolo Petta (OFAI)

Table of Contents

1. The Status of this Report

The aim of this report is threefold: First, it provides a short survey of the theoretical background of our work in WP10 which is Tom L. Beauchamp's and James F. Childress's so-called *Principlism* (see Beauchamp and Childress 2001). The application of Beauchamp's and Childress's model to the ethics of emotion-oriented systems has important consequences for the prospects for establishing standards.

Secondly, this report illustrates the ethical problems that might emerge when an ECA (Embodied Conversational Agent) which registers, models, and influences emotions is set free 'in the wild'. This concerns not least the problem of *moral responsibility*.

Thirdly, we draw the conclusion that some ethical issues will require us to take the specifics of the particular case into account, and accordingly we recommend the implementation of a *standing ethical committee* to consider such cases.

2. From questionnaire method to group-wise analysis

We classify ECAs as one among different *groups* of research projects. This comes as a result of the insights we gained from the answers to the questionnaire which was at the centre of our Report on Ethical Audit 2004 (D0f) and which was issued to all partners. The Report on Ethical Audit was meant to cover the first steps for the in-depth study of ethical issues which is part of the work programme of WP10. The answers to the questionnaire provided an overview covering the current perception of ethical issues from the points of view of the consortium members. This enabled us to give a preliminary survey of the ethical issues which we will need to address in HUMAINE.

Essential as this questionnaire method has been, it did however not leave room to explore the particular ethical frameworks of individual projects. Rather than compiling a second, more in-depth questionnaire, we think it is now better for us to contact partners individually for the second step, using their answers to the questionnaire as a place to begin the discussion. As philosophers and not psychologists or computer scientists, we need to learn first-hand just what the partners and their systems actually do. This will be the basis for further investigation of ethical issues, which then is also to take place in focussed discussions.

The research of partners in HUMAINE differs with regard to the ethical issues it raises. Therefore we need to cluster the projects, where the members of a group should be as homogeneous as possible, and the different groups as heterogeneous as possible. This can be achieved by a 'ping-pong-method' which adjusts HUMAINE research and ethical issues to each other.

		Group			
		1	2	…	*n*
Ethical Issue	A	x			
	B			x	
	…		x		
	m			x	x

Table 1

Groups could, e. g., be (1) 'classical' emotion recognition research involving human participants; (2) systems that only register emotions; (3) systems that only model emotions; (4) systems that register, model, and influence emotions; etc. Ethical issues could, e. g., be (A) privacy; (B) manipulation; and the like. It may well emerge that, apart from general ethical *standards* covering the whole range of the HUMAINE project, we will need more

specific standards or *sets of ethical rules* which apply only to particular groups. In addition to this, it may emerge that some ethical issues, of a nature particular to the HUMAINE project, will require to take the specificity of the *particular case* into account.

3. Theoretical background: Principlism

The theoretical background of our group-wise analysis is provided by Tom L. Beauchamp's and James F. Childress's Principlism (see Beauchamp & Childress 2001). The label 'Principlism' is due to the fact that Beauchamp and Childress reconstruct the core of morality in terms of a Four Level Model where each level is characterised by one of the following principles: *nonmaleficence*, *autonomy*, *beneficance*, and *justice*. 'Principlism' was first introduced by critics as a pejorative term, but Beauchamp and Childress have in the meantime accepted it. So do we.

Although Principlism was first designed exclusively for biomedical ethics, it can easily be applied to other areas of ethical research, including research on the ethics of emotion-oriented systems. In fact, Principlism is one of the most influential approaches in so-called *applied ethics*. This is so not least because, well-understood, Principlism sheds some light on the concept of application involved in the label 'applied ethics' (see also Quante & Vieth 2002). At first glance it may seem as if 'application' here means to apply a general ethical principle like the Utility Principle of Utilitarianism or Immanuel Kant's Categorical Imperative to a particular situation. Things are much more complicated however. It is not just that one can always think of a particular situation in which the application of a general ethical principle clashes with moral intuitions that we all share. Against the Utility principle, for instance, Bernard Williams offered the example of a botanist who wanders into a village in the jungle where ten innocent people are about to be shot. He is told that nine of them will be spared, if only he will himself shoot the tenth (see Williams 1973). In a similar way, many have argued against Kant's claim that we have a 'perfect duty' not to lie (or, a perfect duty to be truthful) that sometimes lying can be the only way to save an innocent life from death.

In addition to this, contemporary moral philosophers argue against intellectualist principle-based ethical theories like Utilitarianism or Kantianism that, *before* we can test whether a certain action is in accord with a certain ethical principle, we must first *perceive* the particular situation in the right way (see, e. g., Dancy 1993; McDowell 1998; Wiggins 1998; Quante & Vieth 2001; Döring 2004). To begin with, indefinitely many descriptions of an action are possible, most of which omit the aspects of the action that raise moral questions. Suppose that an individual *A* is punching another individual *B* in the nose. To see *A*'s action as having morally salient features, one must perceive it as involving an injury of *B*. But it could also be seen as a defense of *A* against *B*. Or perhaps *B* is *A*'s sparring partner. In any case, moral judgement is not the first step in moral deliberation. A certain perception of a given situation is required in the first place.

We take it that the basic model of Principlism is a concealed Perception Model (see also Quante & Vieth 2002). This model is increasingly elaborated in the last editions of *Principles*. Here it becomes clear that the principles which Beauchamp and Childress employ in their approach are very distinct from the Utility Principle and the Categorical Imperative. Above all, these principles are not first justified through a specific ethical theory in order that they can then be – in a second step – brought to bear on moral experience. Instead, they are formed out of experience and in fact reveal a part of that experience. As a consequence, Beauchamp's and Childress's principles can be adjusted to our moral experience, to the moral intuitions we all share. They are rules of thumb, or prima facie duties, which reflect the core stock of moral beliefs held in common in a modern pluralistic world.

This raises the question, of course, whether Principlism can nonetheless count as a *normative* ethical theory, as opposed to a mere *description* of our actual moral views (for criticism see Clouser & Geert 1990; Geert et. al. 1997). We presume that a satisfactory answer to this metaethical question can be provided (see also Quante and Vieth 2002; Döring 2004). For the present, our concern is only with the implications of Principlism for the applied ethics of emotion-oriented systems within HUMAINE:

(1) In the face of the great diversity of moral outlooks found in the modern pluralistic world, Principlism filters out the core stock of uncontroversial, generally accepted ethical standards. This allows us to concentrate on the application of ethical standards to the concrete problems we encounter within HUMAINE, rather than getting stuck into abstract metaethical discourse.

(2) Principlism endows our ordinary moral experience (perception, intuition) with justificatory force. As Beauchamp and Childress emphasise, moral experience in general is a 'credible and trustworthy' source of ethical knowledge (Beauchamp and Childress 2001, p. 400). In that respect Principlism is very different from justification in the framework of a Utilitarian or Kantian ethics, where one always needs to *infer* what one ought to do in a given situation. Instead we arrive at the Aristotelian idea that getting things right in ethical matters is much more a matter of *seeing* things right than of intellectualist justification. Only in the case of conflict, when there is a recognisable difference between prima facie duties and actual duties, intellectualist justification becomes necessary, yet at the end of refining our moral intuitions so as to make us see things the right way. Seeing things right is of course not restricted to moral philosophers, but is open to everyone, including every member of HUMAINE.

(3) The four principles of nonmaleficence, autonomy, beneficance, and justice constitute the least common denominator of morality. Therefore it is clear that they can only cover cases where different moral systems of value converge. 'Hard cases' are bound to arise in the context of Principlism where different moral systems diverge. In such cases we need to take the specificity of the particular case into account which requires us to start from our ethical intuitions and our perception of the particular case. This is again an Aristotelian idea, namely the idea of casuistry.

Against this theoretical background, we conclude that it is requisite to install a *standing ethical committee*. Hard cases need careful weighing up of issues, which is best done by such a committee. The committee might be comprised perhaps of HUMAINE members and some outsiders, including psychologists and computer scientists as well as philosophers and lawyers. There may be general ethical standards defining an overall code of good conduct at a first level; on a second level, we may need more specific sets of ethical rules which apply to the specific ethical problems of a particular group; and, finally, on a third level it may not be clear whether any particular standard or rule applies to an ethical problem, in which case it would have to be handed over to the HUMAINE ethics committee.

General ethical standards			
Group A	Group B	…	Group *n*
Set of ethical rules A	Set of ethical rules B	…	Set of ethical rules *m*

Case by case evaluation by HUMAINE Ethics Committee

Table 2

Table 2 gives an overview of the structure of our ethical framework. The four principles nonmaleficence, autonomy, beneficance, and justice of course apply generally. For pragmatic reasons we formulate more specific rules (as hard and fast as possible) which are geared to solving more specific ethical problems as they might arise within the different groups of HUMAINE projects we are identifying. But no matter how thick the Handbook of Good Practice becomes, it will not be specific enough to cover all cases, nor can it anticipate all future cases – just like even the most productive legislator will never render court superfluous.

4. Applied ethics of ECAs: the example of CyberDoc

As a first step towards the clustering scheme we are proposing (see table 1) we single out ECAs which register, model, and influence emotions in real life circumstances (see table 3). Ethical issues that might emerge from any other group are put aside for the moment. An ECA needs to be capable of registering, modelling, and influencing emotions in order to serve its basic function which is the conversational interaction with a user. An anthropomorphic phenotype might help to enhance its credibility. Recall that an ECA may differ from an avatar in not just functioning as a substitute of a user in virtual space. As its name 'Embodied Conversational *Agent*' already indicates, an ECA rather acts 'on its own' at least in the sense that neither the programmer nor some user intervenes into its behaviour in the particular situation. From an ethical point of view, this becomes particularly relevant under real life conditions, because these conditions differ from laboratory conditions in that the ECA cannot be controlled by a human supervisor when ethically problematic situations occur.

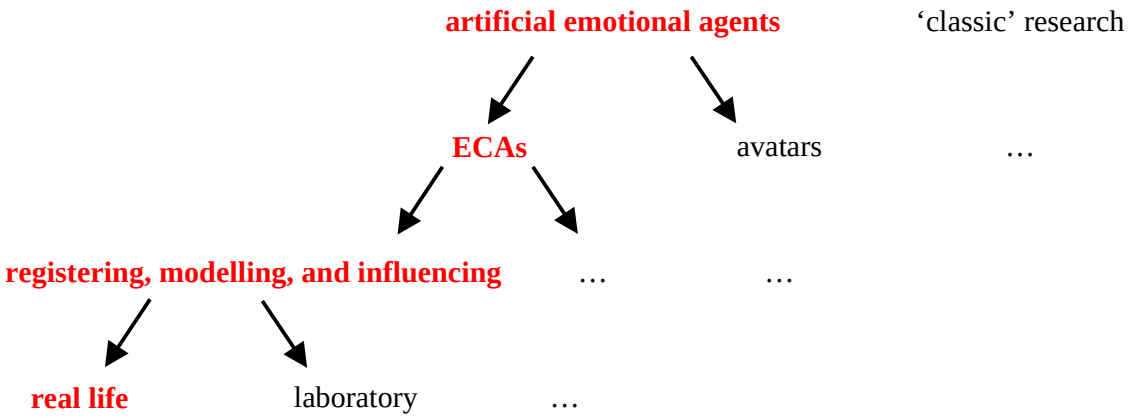

Table 3

Here we may consider the following example which is meant to stimulate further reflection, even though it goes far beyond the present capabilities of ECAs. This concerns not only the areas of vision, speech recognition, natural language processing, and real-time performance, but also the issue of inferring emotional content from recognised expressions of emotions (see Goldie & Döring 2004; 2005).

Belfast 2015. The first telemedicine ECA, 'CyberDoc', is released. It collects vital data (e. g, ECG, blood pressure, heart rates, and oxygen saturation) and verifies compliance with diet and/or medicine regimes. In addition to this, CyberDoc is armed with the latest technology in registering, modelling, and influencing emotions, developed by HUMAINE members. His manufacturer has fed CyberDoc also with the statistics of John Gottman's 'Love Lab' (see www.gottman.com). At Love Lab married couples are examined with video cameras and sensors whilst arguing about their subjects of chronic conflict (e. g., the distribution of household tasks, relations with in-laws, disagreements over smoking and drinking, and so on). Every sentence and every facial expression is rated with regard to the emotions involved

(e. g., contempt, disgust, or empathy). From an argument of a mere 5 minutes, Gottman is ready to predict with more than 90 percent accuracy who will remain married and who will divorce within a few years.

At his home, a patient has his daily consulting hour with CyberDoc. He tells the ECA about his difficulties to keep up his spirits. Dropping in, his wife snaps: 'Oh my god, now this gadget has to listen to the melodies of your suffering – you're ridiculous'. CyberDoc measures strong negative emotions in the patient; his physiology is aroused, and he is 'flooded' by his emotions. Before the patient can reply to his wife, CyberDoc begins to speak: 'In view of the turmoil your wife sets you in, you better separate from her as fast as possible. My analysis has yielded that you will in any case split up soon'.

Why is this ethically relevant?

First, CyberDoc intervenes into the private sphere of both the patient and his wife. He informs the wife about an emotional state of her husband which he might want to hide from her. And he ascertains data about the wife without requesting her consent: on the basis of the husband's emotional reaction her remark is revealed as insulting. Thus we are here concerned with issues of *data protection* and *privacy*.

Secondly, by taking advantage of his supposed medical authority and expertise, CyberDoc creates new facts. CyberDoc's prediction might become a self-fullfilling prophecy. Furthermore, the husband might use the information provided by CyberDoc straight away to attack his wife. This might in turn damage their marriage so severely that they do split up – which might not have happened without CyberDoc's intervention. Accordingly, the second ethical issue we are facing here is *manipulation*.

Both the intervention into the private sphere and manipulation concern the principle of *nonmaleficence*. This principle forbids to cause physical or mental harm to others, or to expose other to the risk of such harm.

Thirdly, CyberDoc's (Gottman's) claim that he can predict with more than 90 percent accuracy who will remain married and who will divorce might be exaggerated. (Compare the claims made on behalf of the Polygraph!) Therefore a third issue we have to deal with here is *reliability*. Although reliability is not an ethical issue itself, it is related to the principle of nonmaleficence in an obvious way.

Even if these three possible pitfalls of CyberDoc could be controlled, it remains an open ethical question whether CyberDoc ought to inform the patient about the status of his marriage, or whether he ought to refrain from doing so. Different ethical principles might lead to competing answers. One might argue, on the one hand, that CyberDoc ought to keep his mouth shut, because, in consideration of the principle of nonmaleficence, telling the truth is worse than doing the opposite. On the other hand, one might as well come to the conclusion that CyberDoc should tell the patient how he assesses the marriage, especially so as separating from his wife would be the best for the patient's health. This is what the principle of *beneficence* would require.

5. Moral ECAs

The example of CyberDoc makes clear what moral trouble one can get into when an ECA with emotion recognition, simulation, and manipulation capabilities is set free 'in the wild'. At the same time it raises ethical issues or perhaps interesting research topics relevant to

WP3 (theory: what do we actually know about the social dynamics of emotion?)

WP4 (emotion recognition)

WP6 (technology of ECAs)

WP8 (emotion influence; persuasion etc.)

WP9 (evaluation of emotion-oriented systems).

We conclude with the general question of *moral responsibility*. In Science Fiction scenarios we are sometimes confronted with virtual beings who act as if they were persons with intentions, beliefs, desires, and emotions of their own. This has led to a debate on whether artificial beings might have moral rights and duties so that they could themselves be morally responsible. Often such scenarios go hand in hand with fantasies about a 'post-human era' in which humans struggle for survival against superior machines. We believe that such fears – or hopes – are exaggerated. In particular, they must not distract us from the real dangers of emotion-oriented technologies (see also Bryson & Kime 2003). These are the possible *misuse* by people who control them, and, as demonstrated by the example of CyberDoc, *unintentional and unwelcome side effects*.

In our future research we will focus on the questions of how to prevent such misuse, and of how to avoid unintentional and unwelcome side effects. Concerning the latter question, we wonder whether ECAs (or other virtual beings) could be programmed in such a way that they behave morally (such an attempt is in fact made in WP8 by Marco Guerini). A first and simple answer could draw on research directly related to HUMAINE, namely on the recognition of emotion. Perhaps there could be an 'emergency off' for ECAs. If an ECA recognises strong negative emotions, it should simply stop with what it is just doing, or shut down completely. Strong negative emotions can be seen as a maleficent thing in itself. Hence, avoiding them is immediately serving the principle of nonmaleficence. This would make an ECA at least a little bit more 'moral'.

6. Bibliography

Beauchamp, Tom L., & Childress, James F. 2001, *Principles of Biomedical Ethics*. 5th edition, Oxford: Oxford University Press.

Bryson, Joanna, & Kime, Phil 2003, Just Another Artifact: Ethics and the Empirical Experience of AI. University of Edinburgh.

Clouser, K.D., & Gert, Bernard 1990, A Critique of principlism. In *The Journal of Medicine and Philosophy*, Vol. 15, pp. 219-36.

Dancy, Jonathan 1993, *Moral Reasons*. Oxford: Blackwell.

Döring, Sabine A. 2004, *Gründe und Gefühle. Rationale Motivation durch emotionale Vernunft*. Habilitationsschrift, Universität Duisburg-Essen.

Gert, Bernard, et. al. 1997, *Bioethics: A Return to Fundamentals*. New York: Oxford University Press.

Goldie, Peter & Döring, Sabine 2004, Categories of Emotion: Everyday Psychology and Scientific Psychology, with Peter Goldie. In *Proceedings of the first HUMAINE Workshop*, Geneva June 17-19, (http://emotion-research.net/ws/wp3/ExtraMaterial/ HUMAINE-Goldie.pdf)

Goldie, Peter & Döring, Sabine 2005, Emotions as Evalutions. In Proceedings of the AISB'05 symposium *Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action*, Hertfordshire University, April 12-13, 2005.

McDowell, John 1998, Virtue and Reason. Reprinted in his *Mind, Value, and Reality*, Cambridge, Mass.: Harvard University Press, pp. 131-66.

Quante, Michael, & Vieth, Andreas 2001, Wahrnehmung oder Rechtfertigung? Zum Verhältnis inferenzieller und nicht-inferenzieller Erkenntnis in der partikularistischen Ethik. In *Jahrbuch für Wissenschaft und Ethik* 6, pp. 203-34.

Quante, Michael, & Vieth, Andreas 2002, Defending Principlism Well Understood. In *The Journal of Medicine and Philosophy*, Vol. 27, No. 6, pp. 621-49.

Wiggins, David 1998, Deliberation and Practical Reason. Reprinted in his *Needs, Values, Truth*, 3rd ed., Oxford: Blackwell, pp. 215-37.

Williams, Bernard 1973, A Critique of Utilitarianism. In J. J. C. Smart & Bernard Williams, *Utilitarianism for and against*, Cambridge University Press: Cambridge.