The Ethical Implications of Non-human Agency in Health Care

Blay Whitby

Centre for Research in Cognitive Science
School of Informatics
University of Sussex
Brighton, BN1 9QJ, UK
blayw@sussex.ac.uk

Abstract

Increased use is being made of, and promises made, for various sorts of deliberately human-like AI technology in health. Computerized Cognitive and Behavioural Therapy (CCBT) is now widely used in the UK and trials are underway elsewhere. In addition to use as a stand-alone therapist, AI technology is increasingly being integrated into therapeutic and care systems. This raises some pressing ethical issues. Studies [e.g.1] have shown that people's reactions to a system performing intimate caring duties are substantially different to their reactions to everyday computer systems. Similarly the work of, for example [2] Breazeal and Scassellati [2], shows that it is possible to manipulate the emotional responses of humans through a robotic system. In health care applications this could be used in ethically desirable or undesirable ways. We need to make a distinction between the two at the design stage of such technologies. Finally there is the problem of moral responsibility for mistakes and clinically undesirable outcomes of AI systems used in health care applications. This problem has often been noticed but never adequately solved. This is an aspect of the technology that merits urgent informed discussion.

The Ethical Implications of Non-Human Agency in Health Care

Ethical Problems in System-Patient Interaction

It seems, at first glance, that the design of health care systems that have more human-like methods of interacting with users is always beneficial. Most of the research in this area is unquestioningly upbeat about the consequences. This is true for example of recently reported research such as [3 and 4]. This paper does not seek to challenge the claims of benefits from these technological developments but rather to point out that there may be a clear possibility of unethical use and undesired outcomes. There are a number of important ethical problems involved in the deployment of non-human agents in health care that require careful consideration. This is a call, therefore, for more attention to be given to the ethical aspects of both the design and the use of this technology.

There is already strong evidence that humans tend to adapt their behaviour to technology to a far greater extent than we know how to adapt technology to humans. The way that this can happen with relatively crude AI technology was originally demonstrated as early as 1966, by ELIZA [5].

ELIZA, and also most state-of-the–art chatbots, output portions of prepared text in response to keywords detected in the user's input. Under favourable conditions, this can *give the illusion* that the system understands and is responding to the user. In fact the system understands nothing. Although he built ELIZA as a joke, Weizenbaum was shocked to observe that people were treating his sytem as if it were capable of genuine conversation [5]. Even users who understand the mechanism behind such systems can be drawn into interacting with the system *as if it were understanding* their input.

In cases where this human tendency to attribute much more understanding than the system actually possesses is limited to a part of an AI research programme, or an amusing game, then no serious ethical issues seem to follow. However the technology is now being directly employed in clinical situations. Computerized Cognitive and Behavioural Therapy (CCBT) is available and approved by the National Institute for Health and Clinical Excellence (NICE) in the UK [6, 7]. An on-line interactive psychotherapy program using this technology called 'Beating the Blues' is in widespread use as a therapeutic intervention in cases of depression or anxiety [7; URL 1].

The approval of NICE is a prima facie indication that the system has proven clinical effectiveness and this paper does not seek directly to challenge the ethical justification for using such systems. A more subtle point is being made.

Beating the Blues shows most clearly that we now have to discuss the ethical behavior, not only of clinical practitioners and support workers, but also of the designers and programmers of AI systems. Designers and programmers have codes of ethics related to their field but these do not at present cover the ethics of clinical interventions. Indeed, all current codes of conduct for computing make the implicit assumption that the users of computer systems are fully competent adults [8]. Nor are they likely, for the foreseeable future, to be extended to cover the more difficult cases of systems intended to be used by more vulnerable individuals.

This process of adaptation will be especially noticeable in cases where AI technology, robots, avatars, and similar devices are used in intimate and caring settings. We should expect this to be especially true in health care situations. For example, empirical studies have shown that humans give more credibility to computer products after they have failed to solve a problem for themselves, or in situations where the human has a strong need for information [1]. This is highly likely to be the case where the system is used to provide health care information. The natural tendency to trust the system will be reinforced by the natural tendency to trust the doctor.

The tendency, demonstrated by the extensive studies of Reeves and Nass [9], of humans to see their interactions with machines in anthropomorphic terms will almost certainly be increased. Interaction designers tend to have mixed feelings about anthropomorphism. Some view it as facilitating good interaction; crucially for present purposes, others take it that it is ethically dubious or misguided to assume it is always beneficial to exploit human emotional and social instincts. For example, Ben Schneiderman describes the human portrayal of a computer as "morally offensive to me" [10]. Nevertheless, in practice, AI technology in health care situations is more likely to be anthropomorphized than previous systems have been, in spite of any worthy intentions of interaction designers.

There is a need to deal with such potential moral offence in the area of health care. It is probably impossible to offer designers a code of conduct that gives clear and comprehensive guidelines on when and when not it is ethical to allow the illusion that an AI system is more human-like than it really is. Sometimes it will be ethically acceptable and even desirable to allow an anthropomorphic illusion. In other cases it will clearly be unethical. One particular unethical use of this illusion is in cases where human responsibility for clinical outcomes is hidden and avoided – "It's the computer's fault."

A further set of ethical issues stems from the tendency of designers unthinkingly to force their view of what constitutes an appropriate interaction on to users. In the field of IT in general there have been many problems caused by this tendency. Some writers, for example Don Norman [11] argue that there is a systematic problem. That is, that computing technology has developed in ways that meet the needs of software companies and most definitely not in ways that meet the needs of users. Even if we are reluctant to concede the full force of Norman's arguments, there would seem to be clear cause for ethical worries about system-patient interactions in health care. Largely unaccountable technical experts may well force their views (both explicit and implicit) of what is appropriate and inappropriate on vulnerable users via this technology.

It is to be hoped that technologists working in the medical area will be forced out of the tendency by the influence of medical professionals. Whether or not this turns out to be the case depends mainly on the quality of collaboration between the two groups. The way the system appears to a user is every bit as much a clinical and ethical matter as it is a technical matter.

Ethical Problems of Responsibility

It is well-understood in medical ethics that there must be a clear trail of responsibility for the care of each and every patient. Unfortunately there is no similar attitude in the field of the ethics of AI and computing in general.

There seems to be little awareness of the need to show a clear trail of responsibility in current AI, HCI, and robotics research and development. This is despite clear warnings having been given [for example, 12, 13]. As has already been remarked, current codes of practice for technologists give no significant guidance in this area. Similarly, the principles of user-centred design – more usually cited than actually followed in current software development practice – are generally based around the notion of creating tools for the user. In this field, by contrast, we might sometimes better describe the goal as the creation of artificial carers for the user. Because of this, there is an urgent need to consider the ethical dimensions of system-patient interaction.

Historically we have been remiss about blaming programmers and designers for the bad consequences of their products. Too often they have been able to avoid responsibility, and the fact that AI systems are inherently less predictable makes this more likely to be possible with non-human agents in health care.

The worst imaginable cases involve an AI system acting 'in loco hominis' in a way that would be clearly unethical if it were done by a human. One way in which this might happen is if the designers of the system follow practices inappropriate for the medical area. Most medical practice is made with the best of intentions, but it is accepted in medical ethics that good intentions alone do not make practice ethically acceptable. This principle applies in at least one precise way to the use of AI and other systems that stand in place of humans.

Deceiving a patient into thinking that they are interacting with a human rather than a machine, for example, often may be ethically acceptable. However it needs to be considered on a case-by-case basis, rather than just waved through. The ethics of particular cases of deliberate deceit is something that demands much of the time of ethics committees – both in health-care contexts and in pure research contexts. Good intentions on the part of AI designers are not sufficient to ensure that what they design is ethically acceptable.

Just as with the ethics of system-patient interaction, a great deal depends on the quality and nature of collaboration between health professionals and computing professionals. At present ethics committees have little or no input into system design, though it is clear that systems can be designed in ethical and highly unethical ways. To take the case of chatbots discussed in the previous section, the illusion that the system is more human-like than is actually the case might often have benefits. On the other hand, such deceit could be used in highly unethical ways. In many cases the system will give the impression of a deeper understanding of, for example, the user's depressive illness than it actually has. In most cases such systems will give the illusion of empathy which might well be expected from a human in a similar role but of which the system is totally incapable. This is not automatically unethical but it certainly deserves further discussion.

It is worth remarking again that the designers of AI systems for patient management or of chatbots used in on-line psychotherapy are not trained in medical ethics. There is no reason for them to have any concern for medical ethics and the ethics of their own field do not, at present, give much guidance. It is to be hoped that this will begin to change and the writer is currently active in drawing up codes of practice for both BCS The Chartered Institute for IT and, at the instigation of the UK funding agency in science and technology, EPSRC, for roboticists [14]. However, contributions to these debates should be from as wide a group as possible. In other fields (such as law and politics) we might reasonably expect decisions with such impacts to be taken in a fully informed and accountable manner including open public debate.

Those who wish to develop and introduce such technologies need to justify their actions to an ethics committee. Such a committee must be able to consider both the design of the technology and the (mainly medical) consequences of its introduction into a clinical setting. It is not ethically acceptable to do this simply because we can.

Discussion

It not a conclusion urged by this paper that work in non-human delivery of health care should be halted or delayed. It is an area that has tremendous potential benefits. It is worth remembering that human delivered health care is *not* always as good as we know that it might be.

However, despite the tremendous usefulness of this sort of technology, the failure to address ethical issues has potentially serious consequences. These include the unintentional limitation of human freedom, the avoidance of moral responsibility for accidents and undesirable outcomes, and the forcing of designers' views as to what is appropriate on especially vulnerable humans.

This paper is a call for a discussion of these and similar ethical issues at an early stage. The issues discussed above repeatedly highlighted the need for high quality collaboration between medical ethicists and system designers so as to prevent unethical practices being tacitly built into the system at an early stage.

We do not need to be quite as critical as Norman [11] of the design of everyday software in order to appreciate that it often inflicts unnecessary anxiety upon users. The promising new technologies using non-human agency in a medical context have the power to affect (usually highly vulnerable) users to a far greater extent.

What is needed is both technically and ethically informed debate on these issues, perhaps with the ultimate goal of being able to provide a code of conduct for designers. It is important to consider these ethical issues with an appropriate urgency.

Acknowledgement

This paper draws heavily upon an earlier unpublished paper written in collaboration with Dr Cavanagh DPhil DClinPsych of the School of Psychology, University of Sussex.

References

[1] Fogg, B. J. and Tseng, H. 1999. Credibility and computing technology. *Communications of the ACM,* 42(5): 39-44.

[2] Breazeal, C. and Scassellati, B. 2002. Robots that imitate humans, *Trends in Cognitive Science*, **6**, pp. 481-487.

[3] Kethuneni, S., August, S. E. and Vales, J. I. 2007. Personal Healthcare Assistant/Companion in Virtual World. *Virtual Healthcare Interaction: Papers from the AAAI Fall Symposium (FS-09-07).*

[4] Elsner, C.H., Berger, T., Wolf, A., Hindricks, G., Mazzi, C. 2003. Healthbot.net: patient education with a

natural speaking robot before catheter ablation: results from 47 patients. *Computers in Cardiology,* 30:669-672.

[5] Weizenbaum 1984. *Computer Power and Human Reasoning* (Pelican edition) pp.188-9.

[6] Proudfoot, J., Swain, S., Widmer, S., Watkins, E., Goldberg, D., Marks, I., Mann, A. and Gray, J.A. 2003. The development and beta-test of a computer-therapy program for anxiety and depression: hurdles and preliminary outcomes. *Computers in Human Behavior,* 19, 277-289.

[7] NICE. 2008. Computerised cognitive behaviour therapy for depression and anxiety: Review of Technology Appraisal 51. *Technology Appraisal 97.* London, UK: NICE.

[8] Whitby, B., 2012, Do You Want a Robot Lover? In Lin, P., Bekey, G. and Abney K. (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

[9] Reeves, B. and Nass, C. 1996. *The Media Equation: how people treat computers, television, and new media like real people and places*. Cambridge, UK: Cambridge University Press.

[10] Don, A., Brennan, S., Laural, B. and Shneiderman, B. 1992. Anthropomorphism: from Eliza to Terminator 2. *Proceedings of the SIGCHI conference on Human factors in computing systems,* Monterey, CA., USA, p 69.

[11] Norman, D. 1999. *The Invisible Computer*, MIT Press, Cambridge, MA.

[12] Picard, R. 1998. *Affective Computing*, Cambridge, A., USA: MIT Press.

[13] Whitby, B. 1988. *Artificial Intelligence: A Handbook of Professionalism,* Ellis Horwood, Chichester.

[14] Baldwin, I., Boden, M. A., Bryson, J. J., Caldwell, D., Dabbs, D., Dautenhahn, K., Duxbury, P., Edwards, L., Grand, A., Grian, H., Kember, S., Kemp, S., Newman, P., O'Dowd, P. J., Parry, V., Pegman, G., Rodden, T., Rose, A., Sorell, T., Wallis, M., West, S., Whitby, B. R. and Winfield, A., 2010, *Principles of Robotics: Regulating Robots in the Real World*, EPSRC report at URL 2

[URL 1] Beating the Blues. Accessed 22.10.2010. http://www.beatingtheblues.co.uk/

[URL2] EPSRC Research Report Accessed 02.03. 2014 http://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx