PsychNology Journal, 2009
Volume 7, Number 1, 49 – 57

Ethical implications of verbal disinhibition with conversational agents

Antonella De Angeli[*♦]

[♦]University of Manchester
(UK)

ABSTRACT

This paper presents a reflection on the ethical implications of conversational agents. The reflection is motivated by recent empirical findings showing that, when interacting in natural language with artificial partners, users tend to indulge in disinhibited behaviour, such as flaming, bullying and sexual harassment. The paper then addresses the question whether conversational agents open any ethical issues and whether this new communication context requires the definition of new moral values and principles or could be addressed by ordinary moral norms.

Keywords: *embodied conversational agents, Internet disinhibition, verbal abuse.*

Paper Received 31/03/2009; received in revised form 29/04/2009; accepted 29/04/2009.

1. Introduction

An on-going debate in computer ethics addresses the uniqueness of the moral dilemmas posed by information technologies (Tavani, 2002). An influential standpoint states that computers generate wholly *new* ethical problems which would have not occurred without technology and for which there are no available analogies in non-computer environment (Maner, 1996). The lack of effective analogies implies that computer ethics requires the definition of new moral values and principles. An alternative standpoint claims that computer ethics transform traditional ethics in complex ways, which anyway can be dealt by within ordinary moral norms (Johnson, 1997). Following this traditionalist approach, three general ethical rules have been proposed for computer-mediated communication in on-line networks (Johnson, 1997).

Cite as:

De Angeli, A. (2009). Ethical implications of verbal disinhibition with conversational agents. *PsychNology Journal, 7*(1), 49 – 57. Retrieved [month] [day], [year], from www.psychnology.org.

[*] Corresponding Author:
Antonella De Angeli
University of Manchester, Manchester Business School, Booth Street West, M
E-mail: antonella.de-angeli@manchester.ac.uk

The first rule claims that on-line communicators must know and abide the norms regulating the forum where the communication takes place. The second rule calls for respecting privacy and property rights of other people. The third one deals more closely with aspects of psychological well-being of the communication partners, stressing the importance of respecting others, not deceiving or harassing them. The similarity between these norms for behaviour on-line and ethical norms for social behaviour in real life is striking.

This paper considers a different context of on-line communication, in which one of the partners is a machine and addresses the question whether this context opens any new ethical issues as compared to human-human interaction. The paper is structured as follows. Section 2 provides a definition of conversational agents alongside an overview of the mainstream research approach dealing with their engineering and an emerging critical rhetorical approach. Section 3 addresses ethical implications of conversational agents and section 4 concludes suggesting ideas for future research.

2. Conversational agents

The term conversational agent describes software which interacts with the user in natural language, via textual input and output or through voice recognition and synthesis. The level of technological complexity varies from pre-scripted chatterbots, which mirror the input of the user through a simple set of transformation rules, to sophisticated multimodal systems, which enrich natural language processing with a number of non-linguistic cues, including hand gestures, facial expressions, and body postures (Abbattista, Catucci, Semeraro, & Zambetta, 2004; Bickmore & Picard, 2005; Cassell, 2000).

Conversational agents are often represented by an anthropomorphic body. A number of embodied conversational agents (ECA's) and talking heads are under development in research centres world-wide and several early prototypes have already entered the Internet. They act as advisors (Berry, Butler & de Rosis, 2005), virtual tutors (Moreno, Klettke, Nibbaragandla, & Graesser, 2002), personal trainers (Bickmore & Picard, 2005) and representatives of major multinational companies (e.g., Ford, Coca-Cola, McDonald and Ikea).

Natural language facilitates anthropomorphic attributions (De Angeli, Gerbino, Cassano, & Petrelli, 2000). Only humans communicate using language and carry on

conversation with one another. Therefore, a talking machine tends to be perceived at a superior level of agency as compared to other machines, and may reach a threshold which subsumes intentionality, sociability, and personality. Anthropomorphic attributions are further strengthened by virtual bodies, which often resemble real humans (Cassell, Bickmore, Campbell, Vilhjalmsson, & Yan, 2001). Overall, conversational interfaces have brought forward an extraordinary change in interaction design: the human metaphor has become the design model (Marakas, Johnson, & Palmer, 2000). ECA's are intentionally designed to be human-like, to show a sense of personality and attitudes, and to involve the user in social relationships.

2.1 Engineering approach

Design and evaluation of ECA's is heavily characterised by a positivistic approach which emphasises desirable consequences of the new interface technology, with little critical analysis of their social implications (Cassell, 2000). The ECA approach found its most influential justification within the field of HCI in the Media Equation Paradigm, a research framework stating that computers are social actors to which users respond as they would do to real people (Reeves & Nass, 1996). This way of thinking deeply influenced the proposition of the Persona Effect claiming that ECA's have the potential for expanding the interaction sphere between humans and computers to encompass social connections and positive emotions leading to more productive task performance (Lester et al., 1997). Despite a recent meta-analysis (Yee, Bailenson & Rickertsen, 2007) has failed to provide empirical support to the Persona Effect, this assumption is still relatively unchallenged in main-stream ECA research which focuses towards the solution of technical issues in the engineering of conversational agents and their evaluation.

2.2 Critical rhetorical approach

A more encompassing research agenda has evolved in the last few years. It stresses the importance of analysing a larger picture of human-agent interaction, one which takes into consideration the full range of human behaviour, including negative outcomes and the possibility of conflicts, disinhibited and antisocial reactions to ECA's (Brahnam & De Angeli, 2008; De Angeli, Brahnam & Wallis, 2005). This critical rhetorical approach shifted the emphasis of ECA research from the engineering of conversational agents to the analysis of their psychological, social and ethical implications.

The approach was motivated by a growing corpus of evidence showing that virtual conversations can at times encourage disinhibited and antisocial behaviour. Verbal abuse and sexual conversations were found to be common elements of anonymous interaction with conversational agents (De Angeli & Brahnam, 2008; Rehm, 2008; Veletsianos, Scharber & Doering, 2008). Independent of the intended focus of the dialogue, interaction setting, and embodiment styles, explicit verbal abuse were common in the user behaviour. An example of verbal abuse is reported in the following *literatim* transcription of a spontaneous interaction between an anonymous web-user and Jabberwacky (www.jabberwacky.com), an on-line chatterbot designed for entertainment.

User: Fuck ypou.
Jabberwacky: I don't think I shall.
User: Fuck you.
Jabberwacky: Don't be superfluously rude.
User: Fuck ur mum.
Jabberwacky: Do you like to eat cake?
User: Off ur mum yeh. [*the interaction continues into a long pornographic talk*]

The user appropriated the chatterbot transforming it into a toy for practicing proposing sex, independently of the fact that the chatterbot did not engage in the sexual conversation and was not explicitly sexually provocative. The conversation closely resembled a context of harassment, which appears to be frequent whenever the agent is represented by an anthropomorphic female body (Brahnam, 2006; De Angeli & Brahnam, 2006; Veletsianos et al., 2008). There is a growing corpus of research indicating that virtual bodies carry with them stereotypical attributions and that user reaction to them is mediated by their physical appearance, such as for example, their gender (Zanbaka, Goolkasian, & Hodges, 2006) ethnicity (Nass, Isbiter & Lee, 2000) and attractiveness (Khan & De Angeli, 2009).

3 Ethical considerations

The occurrence of verbal abuse in human-ECA interaction indicates a need to discuss this topic and explore it more fully and openly. A specific call for action on the ethics of abusing artificial agents, and in particular robots, has been recently put forward (Whitby, 2008). This proposal lays the ground for discussion by proposing

three interdependent ethical issues. First, it raises the question whether it is morally acceptable to treat human-like artefacts in ways that would be considered unacceptable if they would target human beings. Assuming that society considers robot abuse as morally unacceptable, then a new issue is raised as part of the uniqueness debate (Tavani, 2002). It deals with the type of ethical norms which needs to be defined to protect artificial agents, being them unique to this specific context or a direct application of traditional ethics. The final ethical issue is related to design, and considers ways to engineer out the problem of abuse by providing appropriate interaction strategies which constraint its occurrence.

According to Whitby (2008), these questions should be urgently addressed by professional codes of conduct, such as those of the British Computer Society (BCS) and the Association for Computing Machinery (ACM). The argument is justified by the risk that violence towards human-looking artefacts may desensitize the perpetrators, a disputed critique often addressed towards violent video-games (Freier, 2008; Whitby, 2008). The value of an ethical discussion has been strengthened by the application of Christian principles endorsing positive responsibilities, such as “love thy neighbour as thyself” (Dix, 2008). The idea within this perspective is that the ethics of agents can be developed not only by looking at the harm which may come from them, but also at good outcomes, such as the impact of artificial pets on the development of children caring skills. If enhancing positive qualities appeals to our moral sense, then Dix argues that agents that encourage negative behaviours will likewise harm our moral senses. A further elaboration of this way of thinking is that if agents are so anthropomorphic that can be loved by somebody and/or abused by somebody else, their abuse is morally wrong. Thus, designers are encouraged to find new ways to design out unsavoury user behaviours.

A different view-point, claims that there is no pressing need to change current professional codes of conduct because, at the moment, robotics fails to extend the range of social consequences, at least not to the degree that merits special consideration (Thimbleby, 2008). The debate articulates around the appropriateness of applying the concept of abuse to non-sentient agents (Brahnam & De Angeli, 2008). By analysing the issue within the larger context of environmental abuse (those, for example, that lead to global warming) and industrial, personal, and economical abuses of technology (for example, the national financial consequences of not wearing seat belts), the concept of robot abuse is dismissed as sentimentalism. Furthermore, this intellectual position stresses that there are multiple moral layers behind the abuse of

robots. There are some forms of abuse (such as extreme testing in robot war) which have a utilitarian value for engineering development, and as such is morally good.

4 Conclusions

For decades, science fiction writers have envisioned a world in which robots and computers acted like human assistants, virtual companions or artificial slaves. Nowadays, for better or for worse, that world looks closer. As the line between the metaphorical and the literal vanishes, we face uncertainty about how artificial agents will affect our way of interacting with technology, and possibly our social lives. If machines can understand verbal instructions, sense, acquire knowledge, have memory, preferences and personalities many moral and ethical questions are raised. Will they have a sense of self? Who will educate them, guide them, who will they trust? Will there be a time when real and virtual humans are indistinguishable? Who will determine their ethics and morals?

The occurrence of abuse in the interaction with social agents has severe moral, ethical and practical implications. From a moral standpoint, we must reflect on possible effects on individuals, groups, and societies. As we are analysing a dynamic phenomenon which is growing, shaping and constantly changing during the analysis, this reflection must be closely supported by research, which should not only concentrate on the engineering approach but it should move closer to the critical rhetorical approach, proposed in this paper.

The time may not be ready yet for a specific ethics dealing with artificial agents, similar to the growing field of animal ethics or environmental ethics which address moral dilemmas of non-human beings or even inanimate beings. Yet, there is a growing consensus that there is a deontological requirement to initiate a serious reflection within professional bodies, to guide the designers of ECA's in their difficult challenge and possibly to enhance technological development within a value-based approach (Dix, 2008; Freier, 2008; Whitby, 2008). A deontological ethics is of interest to us, if and only if, agent abuse may eventually harm the user. We believe that this risk is intrinsic in the potential of conversational agents to elicit disinhibition and stereotyping.

Stereotypes are widely shared generalisation about members of a specific social groups based on simplified and often derogatory images of out-group members (Fiske

& Taylor, 1991). Stereotypes create a contraposition (us versus them), which may induce and justify anti-social behaviour (e.g., sexism, racism). They are slow and difficult to change, and change requires deeper social and political transformation. We suspect that ECA's may delay this change. Let us take for example sex stereotypes which have been systematically fought by most western nations with positive actions and legal enforcement. ECA's are designed, intentionally or not, with a gender in mind, and more attention is put to the design of attractiveness and realism of female agents. If ECA's encourage gender stereotypes will this impact on real women on-line?

5. References

Abbattista, F., Catucci, G., Semeraro, G., Zambetta, F. (2004). SAMIR: A Smart 3D Assistant on the Web. *PsychNology Journal*, *2*(1), 43-60.

Berry, D. C., Butler, L. T., & de Rosis, F. (2005). Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, *63*(3), 304-327.

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, *12*(2), 293-327.

Brahnam, S. (2006). Gendered bods and bot abuse. In A. De Angeli, S. Brahnam, P. Wallis, A. Dix (Eds.), *Proceedings of the CHI 2006 Workshop: Misuse and abuse of interactive technologies*, (pp. 13-16). Retrieved on October, 3 2009, from agentabuse.org/CHI2006Abuse2.pdf.

Brahnam, S., & De Angeli, A. (2008). Editorial: Special issue on the abuse and misuse of social agents. *Interacting with Computers*, *20*(3), 287-291.

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM*, *43*(4), 70-78.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., & Yan, H. (2001). More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, *14*(1-2), 55-64.

De Angeli, A., & Brahnam, S. (2006-May). Sex stereotypes and conversational agents. Paper presented at the *AVI 2006 workshop on Gender and interaction: Real and virtual women in a male world.* Retrieved on October 3 2009, from: www.informatics.man.ac.uk/~antonella/gender/papers.htm.

De Angeli, A., & Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers*, *20*(3), 302-310.

De Angeli, A., Brahman, S., & Wallis, P. (2005). *Proceedings of Abuse: The darker side of human-computer interaction*, Workshop at Interact 2005. Retrieved on October 3, 2009 from agentabuse.org/Abuse_Workshop_WS5.pdf.

De Angeli, A., Gerbino, W., Cassano, G., & Petrelli, D. (2000-June). From tools to friends: Where is the borderline? Presented at *Proceedings of the UM'99 Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*.

Dix, A. (2008). Response to "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents". *Interacting with Computers*, *20*(3), 334-337.

Fiske, S. T., & Taylor, S. (1991). *Social Cognition*. New York: McGraw-Hill.

Freier, N. G. (2008). Children attribute moral standing to a personified agent. *Proceeding of the SIGCHI conference on human factors in computing systems* CHI 2008 (pp 343-352). New York: ACM Press.

Johnson, D. G. (1997). Ethics online. *Communications of the ACM*, *40*(1), 60-65.

Khan, R., & De Angeli, A. (2009-August). The attractiveness stereotype in the evaluation of embodied conversational agents. Presented at *Interact 2009*.

Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., & Bhogal, R.S. (1997). The persona effect: affective impact of animated pedagogical agents. *Proceeding of CHI97: Human factors in computing systems* (pp. 359-366). New York: ACM Press.

Maner, W. (1996). Unique ethical problems in information technology. *Science and Engineering Ethics*, *2*(2), 137-154.

Marakas, G. M., Johnson, R. D., & Palmer, J. W. (2000). A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model. *International Journal of Human-Computer Studies*, *52*(4), 719-750.

Moreno, K. N., Klettke, B., Nibbaragandla, K., & Graesser, A. C. (2002). Perceived characteristics and pedagogical efficacy of animated conversational agents. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, Lecture Notes in Computer Science 2363, (pp. 963-971), Berlin: Springer Verlag.

Nass, C., Isbister, K., & Lee, E.J. (2000). Truth is beauty: Researching embodied conversational agents. In J. Cassell, Sullivan, J., Prevost, S., Churchill, E. (Eds.), *Embodied Conversational Agents* (pp. 374–402). Cambridge, MA: MIT Press.

Reeves, B., & Nass, C. (1996). *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge: Cambridge University Press.

Rehm, M. (2008). "She is just stupid"--Analyzing user-agent interactions in emotional game situations. *Interacting with Computers*, *20*(3), 311-325.

Tavani, H. T. (2002). The uniqueness debate in computer ethics: What exactly is at issue, and why does it matter? *Ethics and Information Technology*, *4*(1), 37-54.

Thimbleby, H. (2008). Robot ethics? Not yet: A reflection on Whitby's "Sometimes it's hard to be a robot". *Interacting with Computers*, *20*(3), 338-341.

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers*, *20*(3), 292-301.

Whitby, B. (2008). Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers*, *20*(3), 326-333.

Yee, N., Bailenson, J. N., & Rickertsen, K. (2007). A meta-Analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceeding of the SIGCHI conference on human factors in computing systems CHI2007* (pp. 1-10), New York: ACM Press.

Zanbaka, C., Goolkasian, P., & Hodges, L. (2006). Can a virtual cat persuade you?: the role of gender and realism in speaker persuasiveness. In *Proceeding of the SIGCHI conference on human factors in computing systems CHI2006* (pp. 1153-1162), New York: ACM Press.